IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s):	Bentwich et al.	Art Unit:	1631
App. No.:	10/604,944	Examiner:	DeJong, Eric. S.
Conf. No.:	1943	Title:	BIOINFORMATICALLY DETECTABLE GROUP OF NOVEL HIV REGULATORY GENES AND USES THEREOF
Filing Date:	August 28, 2003		

DECLARATION OF DR. ETI MEIRIE UNDER 37 C.F.R. § 1.132

I, Dr. Eti Meirie, hereby declare as follows:

1. I am a Senior Lab Researcher at Rosetta Genomics Ltd., which is the assignee of the above-identified application. A true and correct copy of my *Curriculum Vitae* is attached as Exhibit A.

2. I supervised a project involving the prediction and validation of miRNA precursors and miRNAs in HIV, the details of which are described herein.

3. The nucleic acid "28_1" was sequenced from the BAL isolate strain of HIV (Hwang et al 1992) using methods as described in Bentwich et al 2005. The sequence of 28_1 is presented below aligned against the miRNA precursor from the isolated strain of HIV presented in GeneBank Accession number AF033819 identified as SEQ ID NO: 14 in the above-identified application, with both sequences shown as their DNA equivalent.

```
                                 *
28_1                          AACCTCCAGGGGCAAAT
SEQ ID NO:14 TTAACCCTATAGTGCAGAACATCCAGGGGCAAATGGTACATCAGGCCATATCACCTAGAACTTTAAATGCATGGGTAA
```

4. The highlighted mismatch at position 4 of 28_1 ("C") and position 21 of SEQ ID NO: 14 ("A") occurs at position 747 of the viral genome, within the *gag* gene, which is "C" in the BAL isolate strain and "A" in the strain presented in GeneBank Accession number AF033819.

5. The nucleic acid 28_1 was also cloned two (2) times from the BAL isolate strain of HIV using methods as described in Bentwich et al 2005.

6. The folded hairpin of the miRNA precursor SEQ ID NO: 14 as shown in the above-identified application is shown below.

```
      TATA     GAACA---  C   GGCAA     ACA
TTACCC    GTGCA         TC AGG     ATGGT   T
||||||    |||||         || |||     |||||
AATGGG    TACGT         AG TCC     TACCG   /
      ----     AAATTTCA  A   ACTA-     GAC
```

7. It is my opinion that 28_1 is the BAL isolate strain equivalent of a mature miRNA derived from miRNA precursor SEQ ID NO: 14 through processing by the Dicer enzyme.

8. The HIV genome was analyzed using methods described in Bentwich et al 2005 to identify other potential miRNA precursors obtainable from the *gag* gene (nucleotides 336 to 1838 of the HIV strain presented in GeneBank Accession number AF033819) that may be processed by the Dicer enzyme. Three additional potential hairpins were identified within the *gag* gene; however, each of these potential hairpins received a palgrade score of zero. Based on bioinformatics analysis described in Bentwich et al 2005, each of the three potential hairpins did not pass the minimal stringency criteria of approximately 98% of all known miRNA precursor hairpins that are processed by the Dicer enzyme.

9. Hairpin substrates for the Dicer enzyme are derived from long primary miRNA transcripts (pri-miRNAs) through cleavage by the Drosha enzyme. The pri-miRNAs contain the miRNA-containing hairpin embedded within a longer transcript and are thus surrounded by extra sequences derived from its endogenous flanking genomic sequence. As described in Zeng and Cullen (2005), the sequence of the nucleic acid extensions on either side of the hairpin are mostly unimportant for processing by Drosha.

10. It is my opinion that except for possible extensions at the 5' and/or 3' ends of up to an additional 43 nucleotides, miRNA precursor SEQ ID NO: 14 is the only miRNA precursor obtainable from the HIV *gag* gene.

Dated: August 15, 2006 By: / Eti Meirie /
Dr. Eti Meirie

References

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet. 2005 Jul;37(7):766-70.

Hwang SS, Boyle TJ, Lyerly HK, Cullen BR. Identification of envelope V3 loop as the major determinant of CD4 neutralization sensitivity of HIV-1. Science. 1992 Jul 24;257(5069):535-7.

Zeng Y and Cullen B. Efficient Processing of Primary microRNA Hairpins by Drosha Requires Flanking Nonstructured RNA Sequences. J. Biol. 2005 Jul;280(30): 27595-27603.

Another reference to the HIV BAL isolate strain: "PBMCs infected with viral isolate obtained through AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH cat. # 510"

The HIV strain presented in GeneBank Accession number AF033819 is described in this web page: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=4558520, where the main references is:
Martoglio B, Graf R, Dobberstein B. Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. EMBO J. 1997 Nov 17;16(22):6636-45.

EXHIBIT A

Eti Meiri Curriculum Vitae

NAME Eti Meiri	POSITION TITLE Senior Lab Researcher, Rosetta Genomics		
EDUCATION			
INSTITUTION AND LOCATION	DEGREE (if applicable)	YEAR(s)	FIELD OF STUDY
Tel Aviv University, Tel Aviv, Israel	B.Sc.	1993	Life Science
Hebrew University of Jerusalem, Israel	M.Sc.	1995	Plant Genetics
Weizmann Institute of Science, Rehovot, Israel	PH.D.	2001	Plant Genetics

A. POSITIONS

2002-present Senior lab researcher, Rosetta Genomics, Rehovot, Israel.
2002 Senior researcher at ViroGene biotechnology

B. SELECTED PUBLICATIONS

1. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet.* **2005** Jul;37(7):766-70.
2. MicroRNA Expressio n Detected by Oligonucleotide Microarr ays: System Establishment and Expression Profiling in Human Tissues. Barad O, Meiri E, Avniel A, Aharonov R, Barzilai A, Bentwich I, Einav U, Gilad S, Hurban P, Karov Y, Lobenhofer E, Sharon E, Shiboleth Y.M, Shtutman M, Bentwich Z, and Einat P. *Journal name* **2004,** vol, page-page.
3. Characterization and knock out mutations of three PDI-like genes in *Physcomitrella patens.* in submission. Eti Meiri,Alexander Levitan, Fengying Guo, David A. Christopher, Didier Schaefer, Jean-Pierre Zrÿd, and Avihai Danon. Mol. Genet. Genomics. **2002,** 267:231-240
4. The protein disulfide isomerase-like RB60 is partitioned between stroma and thylakoids in *Chlamydomonas reinhardtii* chloroplasts. Trebitsh T, Meiri E, Ostersetzer O, Adam Z, Danon A. J Biol Chem. **2001**,16;276(7):4564-9.
5. Recombination of engineered defective RNA species produces infective potyvirus inplanta. Gal-On A, Meiri E, Raccah B, Gaba V. J Virol. **1998,** 72(6):5268-70.
6. Simple hand-held devices for the efficient infection of plants with viral-encoding constructs by particle bombardment. . Gal-On A, Meiri E, Elman C, Gray DJ, Gaba V. . J Virol Methods. **1997,** 64(1):103-10.
7. Particle bombardment drastically increases the infectivity of cloned DNA of zucchini yellow mosaic potyvirus. Gal-On A, Meiri E, Huet H, Hua WJ, Raccah B, Gaba V. . J Gen Virol. **1995,** 76 (Pt 12): 3223-7.

C. Abstracts

1. Potyviral recombination using particle bombardment of plants. Gal-On A, Meiri E, Raccah B, Gaba V. Xth nternational congress of Virology, Jerusalem, Israel, 1996.
2. The protein disulfide isomerase-like RB60 is partitioned between stroma and thylakoids in Chlamydomonas reinhardtii chloroplasts. Trebitsh T, Meiri E, Ostersetzer O, Adam Z, Danon A Cellular Implications of Redox Signaling. Abano Terme, Padova, Italy, 2001.
3. Redox Regulated Translation of Chloroplast psbA mRNA in the Light. Trebitsh T, Meiri E, Ostersetzer O, Adam Z, Danon A The Annual Conference of the Israeli Society of Plant Sciences. Rehovot, Israel, 2001.
4. Studying the Role of Protein Disulfide Isomerase-Like Genes in Physcomitrella patens. Meiri E, Dolev D, Pereg, Y, Levitan A, Danon A The Annual Conference of the Israeli Society of Plant Sciences. Rehovot, Israel, 2001.

nature genetics

Identification of hundreds of conserved and nonconserved human microRNAs

Isaac Bentwich[1,2], Amir Avniel[1,2], Yael Karov[1,2], Ranit Aharonov[1,2], Shlomit Gilad[1,2], Omer Barad[1], Adi Barzilai[1], Paz Einat[1], Uri Einav[1], Eti Meiri[1], Eilon Sharon[1], Yael Spector[1] & Zvi Bentwich[1]

MicroRNAs are noncoding RNAs of ~22 nucleotides that suppress translation of target genes by binding to their mRNA and thus have a central role in gene regulation in health and disease[1–5]. To date, 222 human microRNAs have been identified[6], 86 by random cloning and sequencing, 43 by computational approaches and the rest as putative microRNAs homologous to microRNAs in other species. To prove our hypothesis that the total number of microRNAs may be much larger and that several have emerged only in primates, we developed an integrative approach combining bioinformatic predictions with microarray analysis and sequence-directed cloning. Here we report the use of this approach to clone and sequence 89 new human microRNAs (nearly doubling the current number of sequenced human microRNAs), 53 of which are not conserved beyond primates. These findings suggest that the total number of human microRNAs is at least 800.

We developed microRNA discovery tools that detect microRNAs missed by existing methods, which detect only conserved hairpins. Our approach (**Fig. 1a**) comprises the following steps: (i) computationally scanning the entire human genome for hairpin structures; (ii) annotating all hairpins for conserved, repetitive and protein-coding regions; (iii) scoring hairpins by thermodynamic stability and structural features, using a method (PalGrade) that detects a large percentage

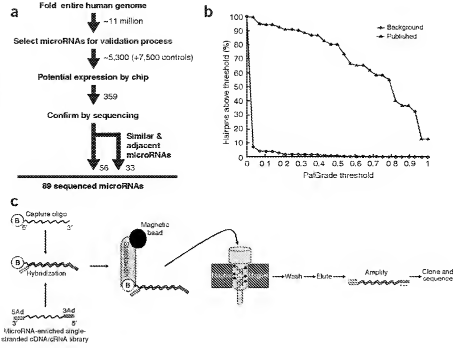


Figure 1 MicroRNA detection and validation. (**a**) The microRNA discovery approach. The initial folding of the entire human genome resulted in ~11 million hairpins. After microarray sampling, 359 microRNAs were subjected to confirmation by sequencing, identifying 89 human microRNAs that did not appear in the microRNA registry: 56 from the main pipeline and 33 additional similar and adjacent microRNAs. (**b**) Hairpin scoring algorithm performance. Percentages of known microRNA hairpins (triangles) and of all genome hairpins (circles) above or equal to different PalGrade thresholds (*x* axis). Hairpins that are not on repetitive or protein-coding regions were considered. The large separation indicates the high sensitivity and specificity (accuracy) of the scoring method. (**c**) The microRNA sequence-directed cloning method. A population of single-stranded molecules derived from the microRNA-enriched library is mixed with the biotinylated capture oligonucleotide. After hybridization, streptavidin bound to magnetic beads is added, and the mixture is loaded into a column mounted on a strong magnet. The column is then washed stringently to remove nonbound or weakly hybridized molecules. The specifically bound molecules are eluted, amplified, cloned and sequenced.

[1]Rosetta Genomics, 10 Plaut Street, Science Park, Rehovot 76706, Israel. [2]These authors contributed equally to this work. Correspondence should be addressed to I.B. (bentwich@rosettagenomics.com).

Published online 19 June 2005; doi:10.1038/ng1590

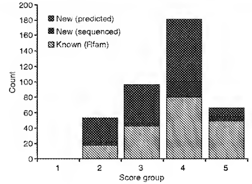


Figure 2 The number of conserved microRNAs in the human genome. Known conserved microRNAs[6] (blue) and new conserved microRNAs (purple), including those validated by us (dashed), for each score group (PalGrade). The five score groups are composed of the following PalGrade ranges: 1, 0 (control group); 2, 0–0.25; 3, 0.25–0.55; 4, 0.55–0.93; and 5, 0.93–1. The number of new microRNAs is the number of hairpins excluding known microRNAs of same score group in the genome, multiplied by the validation success rate in a random sample. The validation success rate is the percentage cloned and sequenced from a sample taken from the group, after deliberate underestimations where success rate was below 5%, divided by the validation success rate of the known microRNAs (76%), to correct for undersampling of tissues (known, similar and adjacent microRNAs were excluded from analysis to avoid positive bias). Because 14% of the known microRNAs were not in the initial group and, hence, are in none of the score groups, these numbers should be divided by 0.86. Thus, the total projected number of conserved microRNAs is 460 (220 + 240).

of known microRNAs while selecting a relatively small portion of all genome hairpins (**Fig. 1b**); (iv) determining the expression of computationally predicted microRNAs by a high-throughput microRNA microarray in several tissues (placenta, testis, thymus, brain and prostate); and (v) validating the sequence of predicted microRNAs that gave high signals on the microarray using a new sequence-directed cloning and sequencing method. This method uses a specific biotinylated capture oligonucleotide, designed for the predicted microRNA to be cloned, to 'fish out' the complementary sequences from the microRNA-enriched libraries, which are then amplified, cloned and sequenced (**Fig. 1c**). This high-throughput approach has enabled us to detect a substantially higher percentage of all existing microRNAs, by dealing effectively with large groups of hairpins that have a relatively high percentage of false positives.

Scanning the entire human genome identified ~11 million hairpins, including 86% of known microRNA precursors. Of all hairpins, 434,239 passed a minimal hairpin score threshold (PalGrade score >0) and were not located on repetitive elements or protein-coding regions. This smaller group, the initial candidate group, retains 86% of known microRNAs, suggesting that systematic scanning would detect 86% of all microRNAs. We then divided all hairpins into conserved and nonconserved hairpins, using a criterion by which 220 of the 222 known microRNAs are conserved. From the initial candidate group we selected ~5,300 predicted microRNA sequences for high-throughput expression validation by microarray[7]. These included randomly sampled hairpins from the following groups: conserved hairpins from different PalGrade score groups (~1,500), nonconserved clustered hairpins from different PalGrade score groups (~800) and nonconserved nonclustered hairpins (~3,000). We also used a control group of ~7,500 hairpins that were not included in the initial candidate group to test various aspects of the prediction approach.

Microarray experiments in placenta, testis, thymus, brain and prostate resulted in 886 candidate microRNAs with significant signals ($P < 0.06$) of at least one of their two predicted mature microRNAs. We subjected 359 of these 886 candidate microRNAs to sequence validation using our new sequence-directed cloning and sequencing

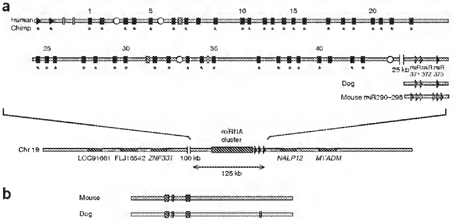


Figure 3 Two new nonconserved microRNA clusters. (**a**) Cluster on chromosome 19 located at positions 58,861,745–58,961,404 (HG17) and comprising 54 microRNA genes grouped into four families on the basis of hairpin sequence similarity (circles, black bars, gray bars and triangles; white bars indicate nongrouped microRNAs). Of these, 43 have been cloned and sequenced (asterisks). The microRNAs of this large family are numbered (for clarity, numbers are shown only for every five microRNAs); numbers match those indicated in **Figure 4a**. The adjacent *mir-371,2,3* cluster is depicted, with its conserved hairpins found in dog and mouse. This cluster is also conserved in rat. (**b**) Cluster on chromosome X located at positions 145,967,859–146,072,859 (HG17). The cluster contains ten cloned and sequenced (asterisks) microRNAs grouped into two families on the basis of hairpin sequence similarity (wide bars and triangles; thin bars indicate nongrouped microRNAs). The microRNAs of the cluster are numbered; numbers match those indicated in **Figure 4b**. Seven precursors in the cluster have homologous hairpins in dog (convergent mapping) and three in mouse (as well as in rat), as indicated by matching colors. The white bars have not mapped to any sequence that can be folded into a hairpin. Neighboring known genes are shown for both clusters. The full list of the microRNA sequences in these clusters is given in **Supplementary Table 1** online.

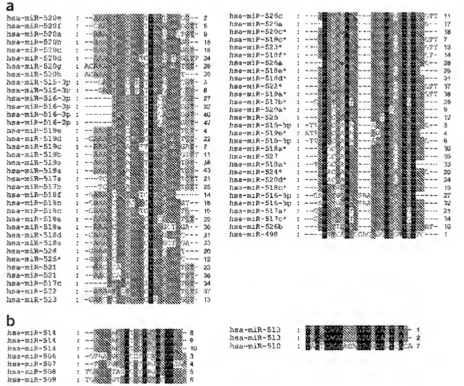


Figure 4 Multiple sequence alignments of cloned mature microRNAs from the two new nonconserved clusters. The multiple alignments were generated by the ClustalW program. Conserved nucleotides are colored as follows: black, 100%; dark gray, 80–99%; and light gray, 60–79%. The miRNA name obtained from the microRNA registry[6] is shown to the left of each sequence. The numbers to the right of each sequence match the numbers indicated in **Figure 3**. (**a**) Multiple sequence alignment of cloned mature microRNAs from the highly related family in the cluster on chromosome 19 derived from the 3′ stem (left column) and 5′ stem (right column) of the precursors. The microRNAs are presented in groups by the 16 distinct seeds they generate (a seed is defined as nucleotides 2–8 of the mature microRNA). Mature microRNAs cloned from the other arm of precursors from which highly expressed microRNAs were cloned are marked with asterisks. (**b**) Multiple sequence alignment of cloned mature microRNAs from the cluster on chromosome X derived from the 3′ stem (left column) and 5′ stem (right column) of the precursors. The microRNAs are presented in groups by the seven distinct seeds they generate.

method (60 of 90 conserved hairpins from different PalGrade score groups; 50 of 72 nonconserved clustered hairpins from different PalGrade score groups; 59 of 161 randomly selected from nonconserved nonclustered hairpins selected from different PalGrade score groups; and 190 of 563 randomly selected from the control group). In some cases, the cloning and sequencing method resulted in sequencing of similar microRNAs that were slightly different in sequence from the microRNA originally sought. We also carried out sequence validation on 69 bioinformatically predicted microRNAs, which were not present on the microarray but are located adjacent to microRNAs that were successfully sequenced, resulting in more new sequences called adjacent microRNAs.

Using this approach, we successfully cloned and sequenced 89 human microRNA genes that do not appear in the microRNA registry[6] (version 5.1; **Supplementary Table 1** and **Supplementary Figs.** 1 and 2 online). Of these, 56 were found through the method's main pipeline (*i.e.*, they were part of the original samples), and 33 are either similar or adjacent microRNAs (these 33 were ignored in calculating success rates). Only 1 of the 89 emerged from the large control group, supporting the distinction between the initial candidate group and the remaining >10 million genomic hairpins (**Fig.** 2). Thirty-two of the 36 conserved microRNAs discovered and sequenced by our approach also score highly on MirScan[8] (**Supplementary Table 1** online). Bioinformatic predictions of conserved microRNAs published after we obtained our results include 32 of these 36 conserved microRNAs (12 appear in the 958 predictions of Berezikov *et al.*[9], 10 in the 129 predictions of Xie *et al.*[10] and 10 in both; of these, 8 were validated by northern-blot or primer extension analysis).

Fifty three of the new microRNAs that we found and sequenced are located in two large nonconserved clusters (24 of these were in the original sample and 29 were found by searching for adjacent microRNAs; **Fig. 3** and **Supplementary Methods** online). One of the clusters, located on chromosome 19 and expressed only in placenta, is the largest microRNA cluster ever reported and comprises 54 new predicted microRNAs, 43 of which we cloned and sequenced (**Fig. 3a** and **Supplementary Tables 1** and 2 online). These 54 microRNAs show similarity to the neighboring *mir-371,2,3* family (**Fig. 3a**) specifically expressed in human embryonic stem cells[11]. Although they are highly similar, they generate 16 distinct 'seeds' (*i.e.*, nucleotides 2–8 of the mature microRNA; **Fig. 4a**). Homology analysis showed that the cluster as a whole is conserved only in chimpanzee and possibly rhesus monkey (*Macaca mulatta*, whose genome is not yet assembled; hence, although all microRNAs are found there, the cluster structure there is not definite) and that none of its microRNA members show any homology to nonprimate genomes. Notably, the adjacent *mir-371,2,3* cluster is conserved in other mammals, although their conservation score was the lowest among the known microRNAs, and one of these microRNAs did not pass the conservation criterion (**Fig. 3a** and ref. 12).

The second cluster is located on chromosome X near the gene *FMR1* and includes ten microRNAs, which are expressed only in testis and all of which we cloned and sequenced (**Fig. 3b** and **Supplementary Table 2** online). These ten microRNAs also form a family of related sequences, which generate seven distinct seeds (**Figs. 3b** and **4b**). The cluster as a whole is conserved only in chimpanzee and possibly rhesus monkey. Seven and four of its members are conserved and clustered in dog and mouse or rat, respectively, although the seven human hairpins converge onto only four hairpins in the dog genome (**Fig. 3b**).

Both clusters differ significantly from known microRNA clusters, all of which are found as a whole in all other mammals (except *mir-371,2,3*, which exists as a whole in dog and partially in rat and mouse; **Fig. 3a**). In addition, none of the microRNAs in the two new clusters has a conservation score passing the criterion discussed above and most cannot be mapped at all to nonprimate genomes. The high cluster-member similarity in both clusters, which are fully conserved only in primates; the convergent mapping of members of the second cluster; and the fact that many of the microRNAs in the first cluster are embedded in long (400–700 nucleotides) sequences that are repeated along the cluster suggest that both clusters evolved through

duplication and mutation events unique to primates. Therefore, we report a new class of microRNAs, nonconserved clustered microRNAs, that are not detected using other microRNA detection approaches and were not taken into consideration in previous estimates of the total number of microRNAs in the genome.

The 89 cloned and sequenced microRNAs that we report bring the total number of human microRNAs to 311, well above a previously stated upper bound of 255 (ref. 4). Moreover, our method allows an estimation of the total number of both conserved and nonconserved microRNA classes. We computed the percentage of true microRNAs comprised by each class using two independent methods: (i) using only informatic data, based on probabilistic arguments, and (ii) calculating the validation success rate in samples from each score group, ignoring similar and adjacent microRNAs to avoid positive bias, and then multiplying the validation success rate by the number of hairpins in the genome belonging to that group (**Supplementary Methods** online). The validation success rate strongly correlated with PalGrade.

The expected number of conserved microRNAs is at least 442 using the probabilistic approach, and at least 460 on the basis of validation success rates, using deliberate underestimations of success rates to account for noisy statistics and correcting for the 14% known microRNAs that were not in the initial group (**Fig. 2** and **Supplementary Methods** online). Estimating the total number of nonconserved microRNAs is more difficult, because this group is new and, hence, less well characterized. We therefore focused on nonconserved hairpins present in a cluster and, as above, obtained two independent estimates for this subgroup. On the basis of validation success rates, these are estimated to be at least 159 nonconserved hairpins. This is probably an underestimate, because our accounting for undersampling of tissues was based on known microRNAs and nonconserved clusters have much higher tissue specificity (**Supplementary Table 2** online). Using the probabilistic approach, we estimate that there are at least 336 nonconserved clustered human microRNAs (**Supplementary Methods** online).

Our results suggest that the world of human microRNAs is larger than initially believed and is not limited to conserved sequences. We estimate that the number of conserved microRNAs is $\sim$400–500 (in accordance with findings of recent studies that used different approaches[9,10]) and that the total number of human microRNAs, including nonconserved clustered microRNAs, is at least 800. These findings support the notion that microRNAs have a central role in the regulation of protein translation throughout the human genome. Our results further indicate that a substantial portion of microRNAs are primate-specific. The fact that the primate-specific clusters described here are specifically expressed in developmental tissues supports the notion that microRNAs may have a key role in the evolutionary process and in the evolved complexity of higher mammals (see also ref. 13).

METHODS

Identifying and scoring candidate microRNA precursors. Step 1: Extracting hairpin structures from the entire genome. We folded the entire human genome using the Vienna package[14] in windows of 1,000 nucleotides with an overlap of 150 nucleotides. All hairpin structures that have at least six base pairs, are at least 55 nucleotides long and have a loop not longer than 20 nucleotides were extracted from the minimum free energy fold of the window (excluding overlapping hairpins). Of the 222 known microRNAs, 14% are missed by this step either because of hairpins that do not fit with our definition (*e.g.*, have a loop that is too large or more than one end-loop) or because of the massive folding in overlapping windows.

Step 2: Assigning each hairpin a stability score. A hairpin is energetically stable if it tends to appear in many folding configurations, which is indicated by the similarity of the minimum free energy graph and the partition function graph provided by the Vienna package for that hairpin. The stability score of the hairpin is 1 minus the mean absolute difference between the two graphs in the hairpin region. The difference is calculated after normalizing the mfe graph so that the mean difference between the two graphs is zero in this region. Thus, scores closer to 1 indicate higher stability.

Step 3: Scoring hairpins. We compared features of known human microRNA precursors with features of a background set of 10,000 randomly selected hairpins found in non–protein-coding regions to identify features that characterize real microRNA precursors relative to background. We used structural features including hairpin length, loop length, stability score, free energy per nucleotide, number of matching base pairs and bulge size, and sequence features including sequence repetitiveness, regular and inverted internal repeats and free energy per nucleotide composition. We constructed an optimal predictor by finding the combination of features that best distinguished between true microRNA precursors and the background set (**Supplementary Methods** online).

Determining hairpin conservation. We divided hairpins into conserved and nonconserved hairpins using the University of California Santa Cruz phastCons[15,16] data. These data contain a measure of evolutionary conservation for each nucleotide in the human genome against the genomes of chimpanzee, mouse, rat, dog, chicken, pufferfish and zebrafish, which is based on a phylogenetic hidden Markov model using best-in-genome pairwise alignment for each species (based on BlastZ), followed by multiZ alignment of the eight genomes[15,16]. We defined a hairpin as conserved if the average phastCons conservation score over the seven species in any 15-nucleotide sequence in the hairpin stem is at least 0.9 (see also ref. 9).

Microarray high-throughput validation. We carried out microarray experiments designed to detect expression of mature microRNAs as previously described[7]. The microarray contains two probes per candidate microRNA gene, one for each predicted mature microRNA. Raw signals vary from a minimal signal of $\sim$400 to a saturated signal of $\sim$65,000. We considered probes with a signal above 2,500 to be positive but not necessarily reliable. To determine the reliability of the signal, we designed a group of 3,000 randomly chosen 35-mers from the human genome and added it to the microarray probes. We observed high correlation ($R^2 = 0.53$) between the probe's maximal signal over the tested tissues and the probe's cytosine (C) content. Only 6% of the background probes with C content below 35% had signals above 2,500 in at least one tissue, whereas more than 70% of the background probes with C content above 35% fulfilled the same condition. Thus, only candidate mature microRNAs with C content below 35% and a signal above 2,500 in at least one tissue passed the microarray high-throughput filtering. Those microRNAs that passed the filtering had a P value of 0.06.

MicroRNA sequence-directed cloning and sequencing. We prepared microRNA enriched libraries as previously described[17] using suitable adaptors. We used RT-PCR amplification with an excess of the reverse primer (1:50 ratio) to produce a cDNA library. We then hybridized biotinylated capture oligonucleotides (22–30 nucleotides long, with biotin at the 5′ end) to an aliquot (5 μl) of the library in TEN buffer. We then added μMACS Streptavidin Microbeads and incubated the reaction for 2 min at the hybridization temperature. We then loaded the mixture onto a magnetized μMACS Streptavidin Kit column and eluted the hybridized single-stranded library molecules by adding 150 μl of water preheated to 80 °C. We recovered the single-stranded cDNA library molecules, amplified them by PCR, ligated them into a pTZ57R/T vector and transformed the ligation products into JM109 bacteria. We identified and sequenced positive colonies (**Supplementary Methods** online).

Determining cluster homology. We compared microRNA precursors with all assembled genomes in the University of California Santa Cruz genome browser (BlastZ analysis[16]). A cluster was considered fully conserved if all its microRNA precursors have homologs that are also clustered. We looked for homologs of individual microRNA cluster members using Blast analysis against the whole-genome sequence data in the National Center for Biotechnology Information Trace databases.

Databases and accession numbers. We submitted new microRNA sequences to the microRNA Registry[6] (Rfam miRNA names are given in **Supplementary Table 1** online). The GEO accession number for microarray data is GSE2708.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank the members of the Rosetta Genomics team for their dedication and contribution.

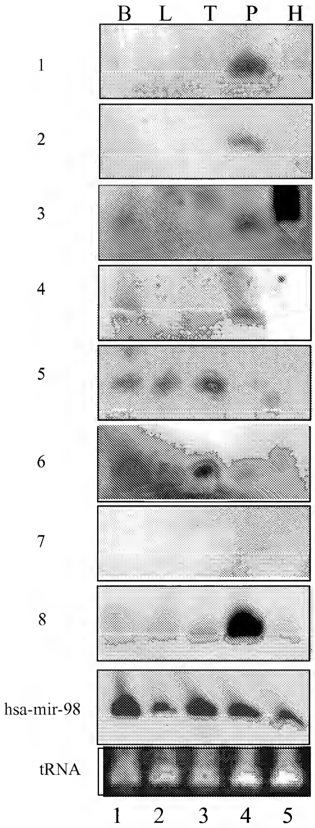
AUTHORS' CONTRIBUTIONS
I.B., A.A., Y.K. and R.A. conceived and designed the methodology of detecting microRNAs and directed the work of the other authors. O.B. designed microarray algorithms, and S.G. conceived the sequencing method. Other authors made significant noninventive contributions: O.B., A.B., P.E., U.E., E.S. and Y.S. developed bioinformatics and algorithmic elements; E.M. and S.G. prepared RNA for microarray analysis and sequenced microRNAs; and Z.B. provided scientific vision and guidance.

COMPETING INTERESTS STATEMENT
The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 24 January; accepted 31 May 2005
Published online at http://www.nature.com/naturegenetics/

1. Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13**, 807–818 (2003).
2. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
3. Johnston, R.J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845–849 (2003).
4. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B. & Bartel, D.P. Vertebrate microRNA genes. *Science* **299**, 1540 (2003).
5. Poy, M.N. *et al.* A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432**, 226–230 (2004).
6. Griffiths-Jones, S. The microRNA registry. *Nucleic Acids Res.* **32**, D109–D111 (2004).
7. Barad, O. *et al.* MicroRNA expression detected by oligonucleotide microarrays: System establishment and expression profiling in human tissues. *Genome Res.* **14**, 2486–2494 (2004).
8. Lim, L.P. *et al.* The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991–1008 (2003).
9. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–24 (2005).
10. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
11. Suh, M. *et al.* Human embryonic stem cells express a unique set of microRNAs. *Dev. Biol.* **270**, 488–498 (2004).
12. Houbaviy, H.B., Murray, M.F. & Sharp, P.A. Embryonic stem cell-specific microRNAs. *Dev. Cell* **5**, 351–358 (2003).
13. Bentwich, I. A postulated role for microRNA in cellular differentiation. *FASEB J.* **19**, 875–879 (2005).
14. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
15. Siepel, A. & Haussler, D. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* **11**, 413–428 (2004).
16. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
17. Elbashir, S.M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).

Supplementary Fig. 1: Expression profiles of novel validated microRNAs (I)

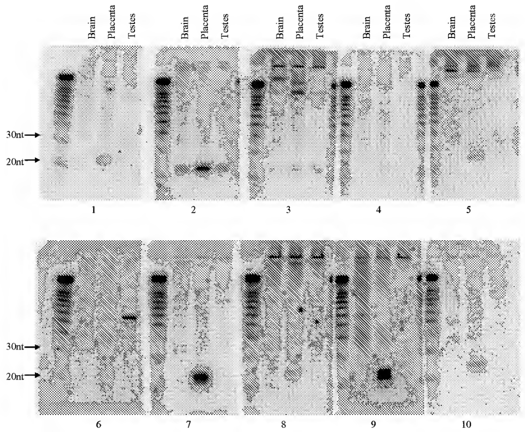


Expression of 8 microRNAs in human brain (B), liver (L), thymus (T), placenta (P) and HeLa cell line (H). Northern blots were made from 40 ug total RNA. Equal loading of the gel before transfer to membrane was monitored by ethidium bromide staining of the tRNA band. The mature microRNA sequences for which results are presented are:

1. hsa-mir-520d
2. hsa-mir-518c
3. hsa-mir-494
4. hsa-mir-432*
5. hsa-mir-497
6. hsa-mir-20b
7. hsa-mir-410
8. hsa-mir-193b

The expression of hsa-mir-98 was also examined for reference. MicroRNAs 2, 5 and 8 also appear in Supplementary Fig. 2 which depicts results for a different set of tissues.

Supplementary Fig. 2: Expression profiles of novel validated microRNAs (II)



Ten Northern blots were created to assess the expression of the microRNAs in three human tissues. 5 ug of total RNA from human brain, placenta, and testis (Ambion) was fractionated by PAGE using a 15% denaturing polyacrylamide gel. The RNA was transferred to positively charged nylon membranes by electroblotting at 200 mA in 0.5X TBE for 2 hours. The Northern blots were dried and incubated overnight in separate hybridization bottles with 10 ml of ULTRAhyb-Oligo (Ambion) and 10^7 cpm of radio-labeled oligonucleotides complementary to the predicted miRNAs. The Northern blots were washed 3 X 10 min at room temperature in 2X SSC, 0.5% SDS and then 1X15 min at 42°C in 2X SSC, 0.5% SDS. Overnight phosphorimaging using the Storm system (Amersham) revealed microRNAs.
The 10 mature microRNAs are:

1- hsa-mir-193b
2- hsa-mir-365
3- hsa-mir-497
4- hsa-mir-512-5p
5- hsa-mir-512-3p
6- hsa-mir-498
7- hsa-mir-515-5p
8- hsa-mir-515-3p
9- hsa-mir-526b
10- hsa-mir-518c

MicroRNAs 1, 3 and 10 also appear in Supplementary Fig. 1 which depicts results for a different set of tissues. Those Northern blots were done by Ambion (Ambion, Inc. 2130 Woodward, Austin, Texas 78744).

Supplementary Table 1: Novel validated microRNAs

The table depicts the list of novel validated microRNAs with their Rfam registry precursor id and chromosomal locations. The mature microRNAs are depicted in red in the draw of the precursor. The tissue in which the microRNA was sequenced is indicated together with the signal on the chip obtained for this microRNA in the relevant tissue, where the number in parenthesis indicates how many of the 185 known microRNAs that were on the chip obtained a higher signal. Also, supplied are number of clones, whether the 5' end was stable among clones, whether the 3' end was variable among clones, whether Northern support was obtained, and for the conserved sequences comparison to related studies: a number indicating their miRscan score (Lim et. al., Genes Dev, 2003) if exists , '+' indicating appearance in the putative miRNAs list of Berezikov et. al (Cell, 2005) and '++' indicating Northern blot support in Berezikov study.

miRNA id & Location	Precursor draw	Clones number	Stable5	Variable3	Northern blot	Sequenced tissues	Chip signal (rank)	Related studies
hsa-mir-488	A CT [illegible] GTTT	1	+	-		Brain	1282 (122)	16.20
Chr1 -	GAGA TCATCT [illegible] CAAACAA \							
173730156-	CTCT AGTAGA GG TCT TTATCG GA AGTTTGTT C							
173730238	C CT T - - A AAAC							
hsa-mir-489	C G CGC TA A							
Chr7 -	GTGGCAG TTGGT GTCGTATGTGTGA CATT CTTGA C							
92757899-	CATCGTC AAT[illegible] GGATT C							
92757982	A A ACA -- T	1	+	-		Brain	1271 (122)	14.14
hsa-mir-490	C -- G CAA T							
Chr7 +	GTTCGACA CATGGA TCTCCAGGT GGT GTT A							
136045199-	CGAGTTG[illegible] CA TAG G							
136045271	[illegible] A CG- A	5	+	+	+	Brain	2922 (105)	18.36, +
hsa-mir-491	TG [illegible] AG	2	+	-		Brain	6001 (100)	15.15
Chr9 +	TTGACTTAGC GGTA[illegible]AGT \							
20706104-	GGTTGGGTCG CCATC TCCCTT GAA GTATTCCTCA A							
20706187	GT T A- C CA							
hsa-mir-511-1	A AC C[illegible] G TA	1	+	-		Placenta	4950 (112)	13.04
Chr10 +	CAAT GAC CCAT [illegible] TAAA T							
17927113-	GTTA CTG GGTA ACAGAAAACGA TGT AGT GTTT T							
17927199	C GT AG TG A - TT							
hsa-mir-511-2	A AC C[illegible] G TA	2	+	-		Placenta	4950 (112)	13.04
Chr10 +	CAAT GAC CCAT [illegible] TAAA T							
18174042-	GTTA CTG GGTA ACAGAAAACGA TGT AGT GTTT T							
18174128	C GT AG TG A - TT							
hsa-mir-146b	T [illegible] GA T	2	-	-		Placenta	2296 (128)	14.58
Chr10 +	CC GGCAC[illegible] GT GC C							
104186259-	GG CCGTGG TCTTGACT AGGTGTCC TA CG T							
104186331	C - C- CG A- A							
hsa-mir-202	T A- G AT	8	-	-		Testes	19369 (29)	15.57, +
Chr10 -	CCGTTCC T[illegible] AGG C							
134949914-	GGCGGGG AAAAGGGTACG ATATGGAGAAA TCC T							
134949989	C GG - GG							
hsa-mir-492	CG[illegible] CCACCA	3	+	+		Placenta	1234 (141)	1.776
Chr12 +	CAT [illegible]GGTG \							
93730667-	GTA CT GACGTCCTGTT AGGACCGC T							
93730739	ACAA A --- AAGAGT							

hsa-mir-493	T CATTC T	4	+	-		Placenta	15619 (92)	12.7, +
Chr14 +	CAGGGCTT [illegible] GT T							
100405158-	GTCCCGGA CGTGTGTCATCTGGAAGTGG CA G							
100405232	C CTTA- C							
hsa-mir-432	[illegible] A TCTATTT	1	+	-		Placenta	7140 (105)	
Chr14 +	CTCCAGG [illegible] TCC \				+			
100420579-	GAGGT[illegible] GGG C							
100420659	[illegible] - TGCATTC	4	+	+		Brain	12038 (88)	5.185
hsa-mir-329-1	CT TT TG TTC T AA							
Chr14 +	GGTAC GAAGAGAGGTT C GGT TGTTTC TT T							
100562875-	CTATG CT[illegible]AAG AG G							
100562954	AC [illegible] C GA	6	-	+		Brain	65518 (1)	12.6
hsa-mir-494	C GT - C - TTTA							
Chr14 +	GATACT GAAGGAGAGGTT CCGTGT TGT TTC TC \							
100565724-	CTATGA TTTTT[illegible] AG T							
100565804	- [illegible] TATT	7	+	+	+	Placenta	6349 (106)	15.64, ++
hsa-mir-495	T T - AT C TAT							
Chr14 +	TGGTACC GAAAAGAAGT GC CCATGTT TTTCG TT \							
100569845-	ACTATGG TTT[illegible] AAAGC AG A							
100569926	C - [illegible] - TGT	5	+	-		Placenta	13172 (98)	16.39
hsa-mir-485	GT TG [illegible] A ATTCA	4	+	+		Brain	2730 (108)	
Chr14 +	TG ACT GAGA[illegible] \							
100591505-	AC TGA T[illegible] GAGC T							
100591585	TG TT [illegible] A - GAAAC	5	+	+		Brain	65518 (1)	15.27, +
hsa-mir-496	T T T GT TTA							
Chr14 +	GGTAC CGAA GGAGGTTG CCATG GTGTTCATT T							
100596672-	TCATG GCTT C[illegible] ATGAGTAG T							
100596750	- T [illegible] TAT	3	-	-		Placenta	3749 (121)	15.81, +
hsa-mir-409	T GG A AC - CA TG							
Chr14 +	TGGTAC CG GAG GGTT CCGAGCAAC TTTG TC \							
100601390-	ACTATG GC T[illegible] AAGC AG G							
100601468	- TT [illegible] -- CA	4	-	-		Brain	43605 (56)	14.56, +
hsa-mir-410	C G TG A A C TTA							
Chr14 +	GGTAC TGAGAA AGGT TCTGTG TG GTTCG TT T							
100602002-	CCATG ACTTT[illegible] AAGC AG T							
100602081	- [illegible] A A - TAA	6	+	-	+	Placenta	9159 (101)	18.1, +
hsa-mir-193b	GT T GT G A A TTAT							
Chr16 +	GTG CTCAGAA CGGG TTTGAGGGC AG TG GT \							
14305325-	TAC GGGT[illegible] AC TA G							
14305407	TG [illegible] A C TTTT	6	+	-	+	Placenta	65518 (1)	14.74, ++
hsa-mir-365-1	AA AC ---- - TTTC							
Chr16 +	G AATGAGGG TTTTGGGGGCA GAT GTG \							
14310651-	T T[illegible] CTA CAC C							
14310723	CG [illegible] T CTTA	6	-	+		Placenta	14379 (95)	11.42, +
hsa-mir-497	[illegible] T -- AC	1	+	-		Placenta	14394 (95)	13.90
Chr17 -	CTCCCGCCC [illegible] AC GGC T				+			
6861968-	GGGGGTGGG TTGT GTGT ACACCAAAC TG CCG G							
6862041	AGCGAGA G C C CA GT							
hsa-mir-365-2	A AC C ----- T T							
Chr17 +	GCAAGAA AATGAGGG TTT AGGGGCA GCTG GTT T							
26926559-	CGTTCTT T[illegible] TGAC CAG C							
26926640	G [illegible] T T	6	-	+		Placenta	14379 (95)	14.57, +

hsa-mir-181d	ATC[illegible] [illegible] GTGA ACT	2	+	-		Placenta	7728 (104)	12.03, ++
Chr19 +	GGTCACA [illegible] GG \				+			
13846714-	TCGGTGT TGTAAGTAG G GGCCACCCAG CC G							
13846849	CAC- G -- A--- GGA							
hsa-mir-512-1	TG [illegible] G	2	-	-	+	Placenta	65518 (1)	
Chr19 +	TCTCAGTC TGG [illegible] [illegible]TG T							
58861745-	GGAGTCAG AC[illegible] AGAC G							
58861828	TA [illegible] AGTA C	2	-	-	+	Placenta	65518 (1)	
hsa-mir-512-2	TG [illegible] G	2	-	-	+	Placenta	65518 (1)	
Chr19 +	CTCAGTC TGG [illegible] [illegible]TG T							
58864230-	GAGTCAG AC[illegible] AGAC G							
58864311	TA [illegible] AGTA C	2	-	-	+	Placenta	65518 (1)	
hsa-mir-498	- CA- [illegible] TATAA	5	-	+		Placenta	65517 (23)	
Chr19 +	CTCAGGCTGTGA [illegible] \				+			
58869284-	GAGTTTGACACT GAAGTTCG CCTCC CGAAAAG C							
58869366	C AGA A TAGGT							
hsa-mir-520e	CT T T G GTTG G							
Chr19 +	TCTC GC GTGACCCTCAAGA GGAAGCA TTTCT TCT A							
58870777-	AGAG TG CATT[illegible] AAAGA AGG A							
58870863	TT T [illegible] A--- A	1	+	-		Placenta	7967 (104)	
hsa-mir-515-1	T C [illegible] GTC	4	+	+	+	Placenta	21415 (90)	
Chr19 +	TCTCA GCAGT A[illegible] [illegible]GTT \							
58874069-	AGAGT TGTCA [illegible] AGACGA T							
58874151	T T [illegible] -- AAG	3	+	-	+	Placenta	65517 (23)	
hsa-mir-519e	T C [illegible] GA	7	-	+		Placenta	8666.5 (104)	
Chr19 +	TCTCA GCAGT A[illegible] TGTTT A							
58875006-	AGAGT TGTCA [illegible] ACAAA A							
58875089	T T [illegible] -- AG	2	+	-		Placenta	65517 (23)	
hsa-mir-520f	T T GTGG AG							
Chr19 +	TCTCAGGC GTGACCCTCTAAAGGGAAGCGCTT CT TC \							
58877225-	AGGGTTTG CA[illegible] GA AG A							
58877311	C C AA-- AA	10	-	+		Placenta	46102 (64)	
hsa-mir-515-2	T C [illegible] GTC	4	+	+	+	Placenta	21415 (90)	
Chr19 +	TCTCA GCAGT A[illegible] [illegible]TT \							
58880075-	AGAGT TGTCA [illegible] [illegible]ACGA T							
58880157	T T [illegible] -- AAG	3	+	-	+	Placenta	65517 (23)	
hsa-mir-519c	C C [illegible] [illegible]G TG	2	+	+		Placenta	1035 (141)	
Chr19 +	TCTCAG CTGTGA C[illegible] TC \							
58881535-	AGAGTT GACAT[illegible] [illegible]AAAGA AG A							
58881621	T [illegible] AA-- AA	5	+	-		Placenta	5621 (108)	
hsa-mir-520a	T C[illegible] G G T	5	+	+		Placenta	12634 (99)	
Chr19 +	CTCAGGCTG GAC [illegible] TT TC \							
58885947-	GAGTTTGGC T[illegible] [illegible]AAAGA AA AG G							
58886031	T [illegible] A G A	3	+	-		Placenta	26791 (82)	
hsa-mir-526b	T C [illegible] GTCT	17	+	+	+	Placenta	49902 (64)	
Chr19 +	TCAGGC GTGA C[illegible] \							
58889459-	AGTCTG CATT [illegible]AAAGA AA G							
58889541	T [illegible] G GAAA	9	+	+		Placenta	24994 (82)	

hsa-mir-							1035 (141)
519b/526c	T T C A GTTG TG	2	-	+		Placenta	
Chr19 +	CA GC GTGA C[illegible] TC \						
58890279-	GT TG CAT[illegible]AAGA AG A						
58890359	T T T A AA-- AA	1	+	-		Placenta	
hsa-mir-525	T C [illegible] A C ATGT	6	+	+		Placenta	65517 (23)
Chr19 +	CTCAAGCTGTGAC [illegible] TT \						
58892599-	GGGTTTGGCATTG [illegible] AA G						
58892683	C A T C A AAAA	1	+	-		Placenta	45564 (65)
hsa-mir-523	TGCT C T GTTG TG	2	+	+		Placenta	1035 (141)
Chr19 +	TCTCA GTGACC[illegible] TC \						
58893451-	AGAGT CATTG[illegible] AAGA AG A						
58893537	TTCC A C AA-- AA	3	-	-		Placenta	57567 (60)
hsa-mir-518f	TGCT C A ---- GT	2	+	+		Placenta	1969 (132)
Chr19 +	TCTCA GTGA C[illegible]T C						
58895081-	AGAGT CATT G[illegible]AAG AGAA T						
58895167	TTCT A C AAA AA	2	-	-		Placenta	12262 (99)
hsa-mir-520b	C GTTG TG						
Chr19 +	CCCTCTA AGGGAAGCGCTTTCT TC \						
58896293-	[illegible]CAAAGA AG A						
58896353	T AA-- AA	9	+	+		Placenta	41880 (70)
hsa-mir-518b	T GC C A TC G TG						
Chr19 +	TCA GCTGTG CCTC AGAGGG AGCGCTT TGTT TC \						
58897803-	AGT TGGCAT [illegible]AA AG A						
58897885	T TT A C -- A AA	5	+	-		Placenta	
hsa-mir-526a-1	CT C A G G T	4	-	-		Placenta	1591 (134)
Chr19 +	CTCAGG GTGA C[illegible] TT CT G						
58901318-	GAGTTT CATT GGAGATTTTCCTTCG GAAAGA AA GA A						
58901402	CT A C G - A						
hsa-mir-520c	T CG GTTG TG	4	-	-		Placenta	1591 (134)
Chr19 +	TCTCAGGC GT TC[illegible] TC \						
58902519-	AGAGTTTG CA [illegible]AAAGA AG A						
58902605	C TT AA-- AA	9	+	+		Placenta	46102 (64)
hsa-mir-518c	T A TC G TG	2	+	-		Placenta	7271 (105)
Chr19 +	GCTGTGAC [illegible] TCTT TC \						
58903816-	TGGCATT[illegible] ACAA AG A						
58903890	T C -- A AA	9	+	+	+	Placenta	62806 (59)
hsa-mir-524	T C A A CTTG A	5	+	+		Placenta	15815 (92)
Chr19 +	TCTCA GCTGTGAC [illegible] TCC \						
58906068-	AGAGT TGGCATTG [illegible]A AGG A						
58906154	T T C C AA-- A	1	+	-		Placenta	40900 (70)
hsa-mir-517a	G T A T GTTGTAT	1	+	-		Placenta	23731 (84)
Chr19 +	TCTCAGGCAGTGAC [illegible] \						
58907334-	AGAGTTTGTCA[illegible]AGA A						
58907420	T C A C AAAGAAA	6	-	+		Placenta	65517 (23)
hsa-mir-519d	T T C C A TC TTT						
Chr19 +	TCCCA GC GTGAC CTC AAAGGGA GCGCTT TGTTTG T						
58908413-	AGGGT TG CAT[illegible]AAAT C						
58908500	T C T A C -- TCT	2	+	-		Placenta	65517 (23)
hsa-mir-521-2	G T T C AA--- C GT						
Chr19 +	TCTCG GC GTGAC CTC AAAGGGAAG TTTT TCTT C						
58911660-	AGAGT TG CAT[illegible] AGAA AGAA T						
58911746	G C T A [illegible] A AA	4	+	+		Placenta	

hsa-mir-520d	G TA C GTTG TA	3	+	+		Placenta	26810 (82)
Chr19 +	TCTCAAGCTGTGA TC GAAAGGGAAG CTTTCT TC \						
58915162-	AGAGTTTGGCATT GG GTTTC[illegible] GAAAGA AG A						
58915248	G TG T AA-- AA	7	+	+	+	Placenta	65517 (23)
hsa-mir-517b	C T A T GTTGTC	1	+	-		Placenta	23731 (84)
Chr19 +	GTGAC [illegible] \						
58916142-	CA[illegible] AGA T						
58916208	T C A C AAAGAA	6	-	-		Placenta	65517 (23)
hsa-mir-520g	T T C - TC GTTGTC						
Chr19 +	TCCCA GC GTGAC CTCTAGAGG AAGCACTT TGTTT \						
58917232-	AGGGT TG CAT[illegible] ACAAA T						
58917321	T C T -- AAAGAG	2	+	+		Placenta	65518 (1)
hsa-mir-516-3	T- T A TAA GTG T	1	+	-		Placenta	38848 (70)
Chr19 +	TCTCA GA GTGAC[illegible] TT T						
58920508-	AGAGT CT CAT[illegible]GAAA AG G						
58920592	TT - T --- GAA T	1	+	-		Placenta	11939 (99)
hsa-mir-526a-2	AG A CT G- G	4	-	-		Placenta	1591 (134)
Chr19 +	GTGACC[illegible] TT A						
58921988-	CATTGGGAGA TTCCT CGT AAGA AG A						
58922052	CT A AC AA A						
hsa-mir-518e	- TG C AG CT GGCT	2	+	+		Placenta	1035 (141)
Chr19 +	TCTCAG GC TGAC [illegible] \						
58924904-	AGAGTT CG AT[illegible]AAAGA A						
58924991	T CA [illegible] AAAGAAA	5	-	+		Placenta	65517 (23)
hsa-mir-518a-1	- - C G GTCT	1	+	-		Placenta	47072 (64)
Chr19 +	TCTCAAGCTGTGA [illegible] TT \						
58926072-	AGAGTTTGGCATT [illegible] GAAAGA AA G						
58926156	A T C G GAAA	3	+	-		Placenta	65517 (23)
hsa-mir-518d	C T C A CT G TG	4	-	-		Placenta	1591 (134)
Chr19 +	TC CA GCTGTGAC [illegible] GTT TC \						
58929943-	AG GT TGGCATTG [illegible] CAA AG A						
58930029	A T C C C- - AA	2	+	-		Placenta	3128 (122)
hsa-mir-516-4	CT A TAA G GT	4	-	+		Placenta	52182 (62)
Chr19 +	TCTCAGG GTGACC [illegible] TTTT \						
58931911-	AGAGTTT CAT[illegible]GAAAGA AAGA G						
58932000	CT G --- A AA	1	+	-		Placenta	11939 (99)
hsa-mir-518a-2	G - C G GTCT	2	-	-		Placenta	49529 (64)
Chr19 +	TCTCAAGCTGTGG [illegible]TCT TT \						
58934399-	AGAGTTTGGCATT [illegible] GAAAGA AA A						
58934485	- T C G GAAA	3	+	-		Placenta	65517 (23)
hsa-mir-517c	AGA T T A T GTTGTC	1	+	-		Placenta	23731 (84)
Chr19 +	GA TCTCAGGCAGTGAC [illegible] \						
58936379-	CT AGAGTTTGTCAT[illegible] TAGA T						
58936473	AAA T T A C AAAGAA	5	-	+		Placenta	65517 (14)
hsa-mir-520h	T T CC - TC GTTGTC						
Chr19 +	TCCCA GC GTGAC TCTAGAGG AAGCACTT TGTTT \						
58937578-	AGGGT TG CAT[illegible] ACAAA T						
58937665	T T -- C -- AAAGAG	2	+	-		Placenta	65518 (1)
hsa-mir-521-1	G T C C AACT GTTG TA						
Chr19 +	TCTCA GC GTGAC CTC AAAGGGAAG TTCT TC \						
58943702-	AGAGT TG CAT[illegible] AAGA AG A						
58943788	G C T A [illegible] AA-- AA	4	+	+		Placenta	

hsa-mir-522	- T TCC [illegible] GTTG TG	2	+	+		Placenta	1035 (141)	
Chr19 +	TCTCAG GC GTG [illegible] TC \							
58946277-	AGAGTT CG CA[illegible] AAAGA AG A							
58946363	T - [illegible] AA-- AA	2	+	-		Placenta	33062 (77)	
hsa-mir-519a-1	T A [illegible] TG	2	+	+		Placenta	1035 (141)	
Chr19 +	CTCAGGC GTGACA[illegible] TC \							
58947463-	GAGTTTG [illegible] AAAGG AG A							
58947547	T A AA-- AA	2	+	-		Placenta	65517 (15)	
hsa-mir-527	[illegible] GTTG TA	1	+	-		Placenta	47072 (64)	
Chr19 +	TCTCAAGCTGTGA [illegible] TC \							
58949084-	AGAGTTTGGCATT GTTTCCCTTCG GAAAGA AG A							
58949168	AAGTG T AA-- AA							
hsa-mir-516-1	C AAA GTTG TG							
Chr19 +	TCTCAGGCTGTGACCTTCT GAGG GAAGCACTTTCT TC \							
58951807-	AGAGTTTGGCAT[illegible] GAAAGA AG A							
58951896	[illegible] --- AA-- AA	1	+	-		Placenta	11939 (99)	
hsa-mir-516-2	GG C AAA GTTG TG							
Chr19 +	TCTCA TTGTGACCTTCT GAGG GAAGCACTTTCT TC \							
58956199-	AGAGT GGCAT[illegible] GAAAGA AG A							
58956288	TT [illegible] --- AA-- AA	1	+	-		Placenta	11939 (99)	
hsa-mir-519a-2	T TCC C A GTTG TG							
Chr19 +	TCTCAGGC GTG CTCTA AGGGA GCGCTTTCT TC \							
58957410-	AGAGTTTG [illegible] AAAGG AG A							
58957496	T [illegible] A AA-- AA	2	+	-		Placenta	65517 (15)	
hsa-mir-499	T [illegible] A [illegible]CTCCT	5	+	+		Prostate	1134 (118)	13.83, ++
Chr20 +	GGC G[illegible] \							
33041867-	TCG CG TCTGAACG CACTACAAGT C							
33041933	T TG A GCACCT							
hsa-mir-500	CCC T- CTA TG AGAG TGTC							
ChrX +	GCTC CTCTC AATCCTTG CC GGTG TGC \							
49476075-	CGAG GAGA[illegible] ACG T							
49476158	AGC [illegible] TAAG	1	+	-		Placenta		14.14
hsa-mir-362	CC - [illegible] AT	4	+	+		Brain	15860 (84)	14.2, +
ChrX +	GCTC CCTCTT GAA[illegible] \							
49476599-	CGAG GGAGAA CTTAGGAACTT ATCCACAC CGTGA T							
49476682	TC A --- AA CT							
hsa-mir-501	T T- [illegible] TTTC	3	+	+		Brain	3492 (104)	8.541, +
ChrX +	GCTC TCCTCTC [illegible] TGC \							
49477366-	CGAG GGGAGAG TTAGGA AC GGG CCCAC ACG T							
49477449	T TC - - - GTA- TAAG							
hsa-mir-502	[illegible] TAG TGGC	2	+	-		Placenta	5403 (110)	3.186, +
ChrX +	[illegible] TGC \							
49482257-	TAGGAACG GGTCCACG ACG T							
49482311	--- TA- TAAC							
hsa-mir-363	GT CA A G----- AG							
ChrX -	GTT CGGGTGGAT CG TGCAATTTT ATG \							
133028929-	CAA [illegible] TAC T							
133029009	[illegible] AGAGGA TA	3	+	+		Placenta	27718 (81)	16.02, ++
hsa-mir-20b	T T [illegible] TTG	1	+	-		Placenta	5827 (107)	15.84
ChrX -	GTCC AG AGTAC [illegible] \				+			
133029352-	TAGG TC TCATG TTCACGGGTAT ATGTC TCAG G							
133029435	T - ACC G ATC TAC							

hsa-mir-450-1 ChrX - 133399897- 133399977	T [illegible] [illegible] CACT GATAC AAACTGT [illegible] \ CTATG TTTGATA ACGTT TACAAGGGTTATAT A T TGT T AAAT	4	+	+		Placenta	22390 (88)	16.23
hsa-mir-450-2 ChrX - 133400073- 133400142	[illegible] [illegible] AAT AAACTAT [illegible]GT \ TTTGATA ACGTT TACAGGGGTTATGTA A CTT T AAT	4	+	+		Placenta	22390 (88)	12.77
hsa-mir-503 ChrX - 133405878- 133405948	A [illegible] T AT TGCCC[illegible] GAGCG \ ATGGGA CGTCGCCTTTGTTA GG GT CTCGT C C T G -- GG	4	+	+		Placenta	58958 (60)	15.84, +
hsa-mir-504 ChrX - 137475392- 137475474	G- - G A- T T GCT CTGT TGGGA[illegible] C GTA T CGG GACA ACCCTTTGGGAC GGACGTGAGG AG CAT C GA T G GA T T	5	-	-		Brain	65518 (1)	17.96, +
hsa-mir-505 ChrX - 138731827- 138731910	AC T G A TCTGC GATGC CCAG GGGGGAGCCAG AAGT TTGATGTT \ CTACG GGTC [illegible]CGA C A- T G C TTTGA	4	+	+	+	Placenta	65517 (23)	15.87, +
hsa-mir-513-1 ChrX - 146000558- 146000623	A G[illegible] TC GAA TGCCT[illegible]GT C ATGGGAAG GTC TCCAC TAAATATA T A TT TT AAA	4	+	-		Testes		
hsa-mir-513-2 ChrX - 146012920- 146012985	A G[illegible] TC GAA TGCCT[illegible]GT C ATGGGAAG GTC TCCAC TAAATATA T A TT TT AAA	4	+	-		Testes		
hsa-mir-506 ChrX - 146017813- 146017877	C A TA A T TGC TTATTCAGGA GGTGT CTTA TAGAT A ATG [illegible] GTTTA A A C -- - T	1	+	-		testes	12279 (39)	
hsa-mir-507 ChrX - 146018055- 146018132	G C A CAT GTC GTGTA TG TTCACTTCAAGA GTGC GCATGT T CACGT AT [illegible] TGTATA A A A C TTT AAG	1	+	-	+	testes	2384 (76)	
hsa-mir-508 ChrX - 146023991- 146024068	G C GT C AAA GTGTA TGC CTACTCCAGAGGGC CA TCATGT C CACAT ATG [illegible]ACA T A A [illegible] T AAA	1	+	-		testes		
hsa-mir-509 ChrX - 146047609- 146047676	C - A G G-- T GTAC CTACT GCAGAC GTG CAATCAT TA A CATG [illegible] AT A A C - G AAA T	2	+	-		testes	4527 (62)	
hsa-mir-510 ChrX - 146059399- 146059472	G TC A- A G[illegible] GTA GTG TG C[illegible] A CAC AT GGTGAG TCTC CA GTTAGTGTG T A GA AA - AA GAT	3	-	+		Testes		
hsa-mir-514-1 ChrX - 146066331- 146066388	- G A T ATA CTACTC TGGA AGTG CAATCA GT A [illegible]AGT TA T T - A T AAT	11	-	-	+	testes		

hsa-mir-514-2	- G A T ATA							
ChrX -	CTACTC TGGA AGTG CAATCA GT A							
146069022-	GATGAG GTCT TCAC GTTAGT TA C							
146069079	T - A T AAT	11	-	-	+	testes		
hsa-mir-514-3	- G A T ATA							
ChrX -	CTACTC TGGA AGTG CAATCA GT A							
146071720-	GATGAG GTCT TCAC GTTAGT TA C							
146071777	T - A T AAT	11	-	-	+	testes		
hsa-mir-452	T AA GT G A TTTGT	2	+	-		Placenta	6153 (106)	
ChrX -	GC AAGCACTTAC CT TTGCAGA GA ACTGAGAC \							
150798668-	CG TTCGTGAATG GA AACGTCT CT TGACTCTG A							
150798752	T AA --- A C TATCA	1	+	-		Placenta	10529 (100)	14.30

Supplementary Table 2: Microarray results for novel validated and known microRNAs
Microarrays were hybridized with 5 µg of Cy3 lcRNA derived from one of the following tissues: placenta, testis, thymus, brain and prostate. Placenta, testis, thymus and brain were hybridized together with 5 µg of Cy5 lcRNA common control derived from a mixture of the above four tissues. For details on probe design see (Barad et. al., Genome Research, 2004). Table presents microarrays Cy3 raw signals which range from background levels of 100-400 to saturation signals of ~65500. Probes with Cy3 signal lower than 1/3 of the common control Cy5 signals were set to background levels to avoid false positives due to interference signals.

mir id	Placenta	Testis	Thymus	Brain	Prostate	chr id	chr start	chr end	strand	Arm
hsa-mir-488	937	332	336	1282	125	1	173730156	173730238	-	PalArm5
hsa-mir-489	1751	392	402	1271	227	7	92757899	92757982	-	PalArm3
hsa-mir-490	419	320	327	2922	491	7	136045199	136045271	+	PalArm3
hsa-mir-491	1135	440	377	6001	408	9	20706104	20706187	+	PalArm5
hsa-mir-511	4950	557	410	523	370	10	17927113	17927199	+	PalArm5
hsa-mir-511	4950	557	410	523	370	10	18174042	18174128	+	PalArm5
hsa-mir-146b	2296	1177	795	1241	501	10	104186259	104186331	+	PalArm5
hsa-mir-202*	302	19369	323	317	116	10	134949914	134959989	-	PalArm5
hsa-mir-492	1234	330	333	351	308	12	93730667	93730739	+	PalArm5
hsa-mir-493	15619	1524	323	852	802	14	100405158	100405232	+	PalArm5
hsa-mir-432	7140	1124	323	7636	1592	14	100420579	100420659	+	PalArm3
hsa-mir-432*	21853	1272	323	12038	481	14	100420579	100420659	+	PalArm5
hsa-mir-329-1	65403	3085	323	65518	1424	14	100562875	100562954	+	PalArm3
hsa-mir-494	6349	973	323	1519	580	14	100565724	100565804	+	PalArm3
hsa-mir-495	13172	914	323	8373	485	14	100569845	100569926	+	PalArm3
hsa-mir-485	65518	304	323	65518	5846	14	100591505	100591585	+	PalArm3
hsa-mir-485*	898	466	335	2730	129	14	100591505	100591585	+	PalArm5
hsa-mir-496	3749	430	353	2021	186	14	100596672	100596750	+	PalArm3
hsa-mir-409	65517	2904	323	43605	4656	14	100601390	100601468	+	PalArm3
hsa-mir-410	9159	467	347	3670	159	14	100602002	100602081	+	PalArm3
hsa-mir-193b	65518	18500	323	62734	7032	16	14305325	14305407	+	PalArm3
hsa-mir-365	14379	1405	323	4270	2176	16	14310651	14310723	+	PalArm3
hsa-mir-497	14394	36876	323	29507	50988	17	6861968	6862041	-	PalArm5
hsa-mir-365	14379	1405	323	4270	2176	17	26926559	26926640	-	PalArm3
hsa-mir-181d	7728	382	699	30518	1346	19	13846714	13846849	+	PalArm5
hsa-mir-512-3p	65518	304	323	317	129	19	58861745	58861828	+	PalArm3
hsa-mir-512-5p	65518	304	323	317	103	19	58861745	58861828	+	PalArm5
hsa-mir-512-3p	65518	304	323	317	129	19	58864230	58864311	+	PalArm3
hsa-mir-512-5p	65518	304	323	317	103	19	58864230	58864311	+	PalArm5
hsa-mir-498	65517	304	323	317	128	19	58869284	58869366	+	PalArm5
hsa-mir-520e	7967	326	338	340	573	19	58870777	58870863	+	PalArm3
hsa-mir-515-3p	65517	304	323	317	150	19	58874069	58874151	+	PalArm3
hsa-mir-515-5p	21415	304	323	369	89	19	58874069	58874151	+	PalArm5
hsa-mir-519e	65517	304	323	317	956	19	58875006	58875089	+	PalArm3
hsa-mir-519e*	8667	304	323	317	116	19	58875006	58875089	+	PalArm5
hsa-mir-520f	46102	304	323	317	3022	19	58877225	58877311	+	PalArm3
hsa-mir-515-3p	65517	304	323	317	150	19	58880075	58880157	+	PalArm3
hsa-mir-515-5p	21415	304	323	369	89	19	58880075	58880157	+	PalArm5
hsa-mir-519c	5621	315	351	334	416	19	58881535	58881621	+	PalArm3
hsa-mir-519c*	1035	314	335	339	95	19	58881535	58881621	+	PalArm5
hsa-mir-520a	26791	304	323	317	216	19	58885947	58886031	+	PalArm3
hsa-mir-520a*	12634	304	323	317	455	19	58885947	58886031	+	PalArm5

hsa-mir-526b*	24994	304	323	317	1482	19	58889459	58889541	+	PalArm3
hsa-mir-526b	49902	304	323	317	5231	19	58889459	58889541	+	PalArm5
hsa-mir-526c	1035	314	335	339	95	19	58890279	58890359	+	PalArm5
hsa-mir-525*	45564	304	323	317	306	19	58892599	58892683	+	PalArm3
hsa-mir-525	65517	304	323	317	247	19	58892599	58892683	+	PalArm5
hsa-mir-523	57567	304	323	317	203	19	58893451	58893537	+	PalArm3
hsa-mir-523*	1035	314	335	339	95	19	58893451	58893537	+	PalArm5
hsa-mir-518f	12263	304	323	317	98	19	58895081	58895167	+	PalArm3
hsa-mir-518f*	1970	308	329	340	396	19	58895081	58895167	+	PalArm5
hsa-mir-520b	41881	304	323	317	2714	19	58896293	58896353	+	PalArm3
hsa-mir-526a	1592	309	332	332	304	19	58901318	58901402	+	PalArm5
hsa-mir-520c	46102	304	323	317	3022	19	58902519	58902605	+	PalArm3
hsa-mir-520c*	1592	309	332	332	304	19	58902519	58902605	+	PalArm5
hsa-mir-518c	62806	304	323	317	98	19	58903816	58903890	+	PalArm3
hsa-mir-518c*	7271	304	323	317	1852	19	58903816	58903890	+	PalArm5
hsa-mir-524	40901	304	323	317	329	19	58906068	58906154	+	PalArm3
hsa-mir-524*	15815	304	323	317	249	19	58906068	58906154	+	PalArm5
hsa-mir-517a	65517	304	323	317	7514	19	58907334	58907420	+	PalArm3
hsa-mir-517a*	23731	304	323	317	127	19	58907334	58907420	+	PalArm5
hsa-mir-519d	65517	304	323	317	5946	19	58908413	58908500	+	PalArm3
hsa-mir-520d	65517	304	323	317	92	19	58915162	58915248	+	PalArm3
hsa-mir-520d*	26811	304	323	317	4369	19	58915162	58915248	+	PalArm5
hsa-mir-517b	65517	304	323	317	7514	19	58916142	58916208	+	PalArm3
hsa-mir-517b*	23731	304	323	317	127	19	58916142	58916208	+	PalArm5
hsa-mir-520g	65518	304	323	317	538	19	58917232	58917321	+	PalArm3
hsa-mir-516-3p	11940	304	323	363	402	19	58920508	58920592	+	PalArm3
hsa-mir-516-5p	52183	306	329	323	100	19	58920508	58920592	+	PalArm5
hsa-mir-526a	1592	309	332	332	304	19	58921988	58922052	+	PalArm5
hsa-mir-518e	65517	304	323	317	247	19	58924904	58924991	+	PalArm3
hsa-mir-518e*	1035	314	335	339	95	19	58924904	58924991	+	PalArm5
hsa-mir-518a	65517	304	323	317	197	19	58926072	58926156	+	PalArm3
hsa-mir-518a*	49530	304	323	317	2558	19	58926072	58926156	+	PalArm5
hsa-mir-518d	3129	323	346	358	128	19	58929943	58930029	+	PalArm3
hsa-mir-518d*	1592	309	332	332	304	19	58929943	58930029	+	PalArm5
hsa-mir-516-3p	11940	304	323	363	402	19	58931911	58932000	+	PalArm3
hsa-mir-516-5p	52183	306	329	323	100	19	58931911	58932000	+	PalArm5
hsa-mir-518a	65517	304	323	317	197	19	58934399	58934485	+	PalArm3
hsa-mir-518a*	49530	304	323	317	2558	19	58934399	58934485	+	PalArm5
hsa-mir-517c	65518	304	323	317	3816	19	58936379	58936473	+	PalArm3
hsa-mir-517c*	23731	304	323	317	127	19	58936379	58936473	+	PalArm5
hsa-mir-520h	65518	304	323	317	538	19	58937578	58937665	+	PalArm3
hsa-mir-522	33062	304	323	317	110	19	58946277	58946363	+	PalArm3
hsa-mir-522*	1035	314	335	339	95	19	58946277	58946363	+	PalArm5
hsa-mir-519a	65518	304	323	317	877	19	58947463	58947547	+	PalArm3
hsa-mir-519a*	1035	314	335	339	95	19	58947463	58947547	+	PalArm5
hsa-mir-527	47072	304	323	317	2088	19	58949084	58949168	+	PalArm5
hsa-mir-516-3p	11940	304	323	363	402	19	58951807	58951896	+	PalArm3
hsa-mir-516-3p	11940	304	323	363	402	19	58956199	58956288	+	PalArm3
hsa-mir-519a	65518	304	323	317	877	19	58957410	58957496	+	PalArm3
hsa-mir-499	383	346	333	622	1134	20	33041867	33041933	+	PalArm5
hsa-mir-362	65518	2842	323	15860	4950	X	49476599	49476682	+	PalArm5
hsa-mir-501	1349	448	599	3492	381	X	49477366	49477449	+	PalArm5

hsa-mir-502	5403	534	444	959	422	X	49482257	49482311	+	PalArm5
hsa-mir-363	27718	1023	323	10664	12520	X	133028929	133029009	-	PalArm3
hsa-mir-20b	5827	411	380	2653	6074	X	133029352	133029435	-	PalArm5
hsa-mir-450	22390	3180	323	317	1164	X	133399897	133399977	-	PalArm5
hsa-mir-450	22390	3180	323	317	1164	X	133400073	133400142	-	PalArm5
hsa-mir-503	58958	1201	323	317	339	X	133405878	133405948	-	PalArm5
hsa-mir-504	4817	4874	323	65518	1396	X	137475392	137475474	-	PalArm5
hsa-mir-505	65517	20767	323	65518	16042	X	138731827	138731910	-	PalArm3
hsa-mir-506	302	12279	323	317	229	X	146017813	146017877	-	PalArm3
hsa-mir-507	302	2394	323	317	94	X	146018055	146018132	-	PalArm3
hsa-mir-509	302	4527	323	317	116	X	146047609	146047676	-	PalArm3
hsa-mir-452-3p	10529	360	344	414	823	X	150798668	150798752	-	PalArm3
hsa-mir-452-5p	6153	322	340	335	823	X	150798668	150798752	-	PalArm5

mir id	**Placenta**	**Testis**	**Thymus**	**Brain**	**Prostate**
hsa-let-7a	34407	15978	323	39777	22949
hsa-let-7b	51379	65518	323	65518	33188
hsa-let-7c	5800	8603	323	29521	20840
hsa-let-7d	5578	4218	323	16714	4609
hsa-let-7e	6066	3378	323	9809	4234
hsa-let-7f	3825	2702	323	10363	9292
hsa-let-7g	332	308	331	332	105
hsa-let-7i	21166	14837	323	65518	22695
hsa-mir-1	307	306	327	323	110
hsa-mir-100	65518	5888	323	44806	8359
hsa-mir-101	2492	759	342	1058	901
hsa-mir-103	34438	4306	323	23689	5743
hsa-mir-105	4796	466	335	30286	344
hsa-mir-106a	65517	304	323	36142	19661
hsa-mir-106b	22447	304	1347	2907	1888
hsa-mir-107	24481	4314	323	17766	4175
hsa-mir-10a	23082	3704	323	1819	21995
hsa-mir-10b	302	24198	323	317	7829
hsa-mir-122a	332	318	331	357	114
hsa-mir-124a	302	304	1532	65518	259
hsa-mir-125a	65517	65517	323	65518	29764
hsa-mir-125b	65518	65518	323	65518	65233
hsa-mir-126	65517	40553	323	65518	65233
hsa-mir-126*	8752	3756	323	2804	3472
hsa-mir-127	63499	10079	323	65518	2355
hsa-mir-128a	3856	1409	1256	65518	2159
hsa-mir-128b	14494	3974	323	65518	5346
hsa-miR-129	1814	1259	464	65518	3425
hsa-mir-130a	26975	1253	323	1327	4660
hsa-mir-130b	1547	321	378	474	190
hsa-mir-132	12833	5429	323	65518	10719
hsa-mir-133a	65518	12003	323	65518	65233
hsa-miR-133b	65517	10769	323	65518	65233
hsa-mir-134	4730	1568	323	6163	514
hsa-mir-135a	1595	1066	344	1210	2152
hsa-miR-135b	14962	1217	323	2038	301

hsa-mir-136	10045	959	323	493	300
hsa-mir-137	621	315	348	1449	354
hsa-mir-138	23885	304	749	65518	719
hsa-mir-139	65518	13141	323	65518	22213
hsa-mir-140	38024	14602	323	19448	19529
hsa-mir-141	63447	304	323	317	6888
hsa-mir-142-3p	65517	304	37848	317	26775
hsa-mir-142-5p	393	333	353	356	127
hsa-mir-143	56256	9637	323	5647	65233
hsa-mir-144	942	350	337	348	110
hsa-mir-145	65517	65518	323	65220	65233
hsa-mir-146	779	377	796	650	446
hsa-mir-147	348	305	333	321	91
hsa-mir-148	35998	21437	323	16814	34644
hsa-miR-148b	65518	8001	323	65518	9909
hsa-mir-149	65517	304	323	65518	5868
hsa-mir-150	28724	12247	323	30799	9418
hsa-miR-151	65517	7548	323	65091	13873
hsa-mir-152	65518	11591	323	26345	51479
hsa-mir-153	369	390	341	2770	554
hsa-mir-154	13926	2796	323	2041	426
hsa-mir-155	302	304	13331	317	1781
hsa-mir-15a	65517	11374	323	29098	35946
hsa-mir-15b	65517	28255	323	61351	9677
hsa-mir-16	65110	7289	323	20739	14915
hsa-mir-17-3p	924	324	378	349	136
hsa-mir-17-5p	65518	304	6516	19761	16301
hsa-mir-18	4710	304	1377	947	472
hsa-mir-181a	65517	304	323	65518	11617
hsa-mir-181b	45856	304	2420	65518	2878
hsa-mir-181c	13966	522	500	18896	1324
hsa-mir-182	22319	304	323	8353	14037
hsa-mir-182*	302	304	323	317	151
hsa-mir-183	4254	555	706	1387	3665
hsa-mir-184	413	355	331	382	100
hsa-mir-185	2400	442	357	1866	274
hsa-mir-186	6935	1000	421	2203	2310
hsa-mir-187	524	559	337	1476	3209
hsa-mir-188	2283	616	341	589	370
hsa-mir-189	14462	1522	323	8084	7300
hsa-mir-190	336	308	333	328	109
hsa-mir-191	65517	17362	323	65518	19397
hsa-mir-192	8716	599	323	1412	809
hsa-mir-193	4308	1609	323	552	731
hsa-mir-194	52307	304	323	20414	10228
hsa-mir-195	3059	5414	323	4448	6120
hsa-mir-196	1500	443	367	336	1393
hsa-mir-197	41922	10638	323	65518	8523
hsa-mir-198	1341	451	339	609	1043
hsa-mir-199a	65517	65517	323	317	65233
hsa-mir-199a*	65517	45113	323	317	38623
hsa-mir-199b	65517	22715	323	317	65223

hsa-mir-19a	1245	410	652	469	314
hsa-mir-19b	44004	2230	18039	13706	27304
hsa-mir-20	5402	420	683	1383	966
hsa-mir-200a	434	312	333	337	465
hsa-mir-200b	586	383	340	431	3509
hsa-mir-200c	65517	304	323	317	10127
hsa-mir-203	2257	345	345	373	634
hsa-mir-204	44061	18811	323	65518	19815
hsa-mir-205	65517	304	323	317	15078
hsa-mir-206	314	309	331	322	271
hsa-mir-208	308	311	330	323	102
hsa-mir-21	65517	56319	44246	60898	65233
hsa-mir-210	30571	304	5532	6363	1333
hsa-mir-211	11512	4064	323	43524	3842
hsa-mir-212	2164	2203	1886	39108	3038
hsa-mir-213	28576	1139	1841	41102	892
hsa-mir-214	65518	65518	323	317	20534
hsa-mir-215	328	312	332	326	102
hsa-mir-216	362	316	376	399	117
hsa-mir-217	327	308	333	379	113
hsa-mir-218	13809	1013	336	11219	5927
hsa-mir-219	432	325	418	8751	142
hsa-mir-22	65518	29110	323	65518	47250
hsa-mir-220	718	375	385	561	286
hsa-mir-221	65518	15883	323	65518	65233
hsa-mir-222	65518	304	323	65518	51527
hsa-mir-223	65517	9609	323	32507	8155
hsa-mir-224	42729	304	323	317	2649
hsa-mir-23a	65517	25778	323	46367	24526
hsa-mir-23b	65517	23368	323	65518	60132
hsa-mir-24	65518	27058	323	53946	37761
hsa-mir-25	65517	26210	323	65518	25097
hsa-mir-26a	65518	40972	323	46059	49881
hsa-mir-26b	20792	4000	323	6213	20451
hsa-mir-27a	65517	65517	323	65508	65233
hsa-mir-27b	65517	65517	323	65518	65233
hsa-mir-28	3128	1415	323	3787	2445
hsa-mir-296	28936	304	323	33381	25089
hsa-mir-299	65518	12104	323	51174	4222
hsa-mir-29a	38248	10115	323	20891	14528
hsa-mir-29b	44087	5436	323	8613	12737
hsa-mir-29c	3537	1349	415	2016	1608
hsa-mir-301	23046	304	323	3166	2825
hsa-mir-302	329	308	330	425	97
hsa-miR-302a	329	308	330	425	97
hsa-miR-302a*	311	309	334	323	97
hsa-miR-302b	370	313	326	371	99
hsa-miR-302b*	341	310	336	326	95
hsa-miR-302c	377	309	338	407	99
hsa-miR-302c*	383	354	368	410	164
hsa-miR-302d	403	304	328	360	99
hsa-mir-30a	65517	12456	323	65518	32041

hsa-mir-30a*	65517	24899	323	65518	65233
hsa-mir-30b	65518	304	323	57387	37418
hsa-mir-30c	65518	29857	323	65518	64861
hsa-mir-30d	65517	28999	323	65518	15596
hsa-mir-30e	65517	304	323	44620	50173
hsa-mir-31	3978	992	323	6583	3402
hsa-mir-32	1358	398	342	399	335
hsa-mir-320	65518	2613	323	14277	3670
hsa-mir-321	65518	304	323	12070	65233
hsa-miR-323	59256	2270	323	65518	2609
hsa-miR-324-3p	23582	4045	323	38306	3994
hsa-miR-324-5p	65517	304	323	65454	22712
hsa-miR-326	36061	6426	323	65518	8800
hsa-miR-328	65518	20890	323	65518	15963
hsa-mir-33	2661	3108	323	1723	2223
hsa-miR-330	2079	324	529	16932	554
hsa-miR-331	51723	12933	323	65518	8943
hsa-miR-335	464	350	359	390	150
hsa-miR-337	36827	5003	323	317	1219
hsa-miR-338	791	444	395	12681	931
hsa-miR-339	65518	63268	323	65518	16768
hsa-miR-340	5470	872	323	19968	2882
hsa-miR-342	64682	10765	323	65518	14624
hsa-mir-34a	62936	12585	323	54785	12889
hsa-mir-34b	1283	447	324	649	140
hsa-mir-34c	1288	499	334	573	134
hsa-miR-367	303	305	328	317	86
hsa-miR-368	687	347	368	361	107
hsa-miR-369	541	332	339	499	106
hsa-miR-370	4502	1400	323	23542	329
hsa-miR-371	7579	350	344	366	155
hsa-miR-372	65517	304	323	317	144
hsa-miR-373	5268	351	353	346	140
hsa-miR-373*	625	389	391	564	237
hsa-miR-374	400	334	337	1921	135
hsa-mir-7	2334	401	553	4299	591
hsa-mir-9	847	304	1021	32931	720
hsa-mir-9*	316	310	332	439	91
hsa-mir-92	65518	65517	65518	65518	7684
hsa-mir-93	65517	304	323	40489	7943
hsa-mir-95	589	850	436	9170	900
hsa-mir-96	9073	696	565	655	5691
hsa-mir-98	359	333	361	585	174
hsa-mir-99a	22079	26900	323	60079	57772
hsa-mir-99b	65518	11694	323	65518	19622

Supplementary Methods

Identifying Candidate MicroRNA Precursors: Step 3 - Scoring Hairpins

The hairpin structure characteristics that were selected are: a) hairpin length, b) loop length, c) stability score (as defined above), d) free energy per nucleotide, e) matching base pairs - the maximal number of paired nucleotides in a sliding window of 22nts length starting at positions 31 to 27nts from the center of the hairpin loop on the 5' arm, and f) bulge size – the minimal largest number of sequential unpaired nucleotides on the sliding window as defined for the previous feature. The hairpin sequence characteristics that were selected were: a) sequence repetitiveness – the maximal number of times the two most abundant dinucleotides (AA, AT, etc.) appear in any subsequence of 22nts length, b) regular internal repeat - the length of the longest non-overlapping repeated sequence in the hairpin, c) inverted internal repeat - the length of the longest non-overlapping sequence repeated in opposite orientation within the hairpin, d) free energy/nucleotide composition (Kcal/mol) - the Z score of the minimum free energy of the hairpin when compared to the distribution of free energy values of random sequences in the genome with similar lengths and nucleotide distributions (see also [1]), and e) GC content – the GC content in the hairpin arm with the lower GC content.

The optimal predictor is constructed by finding the combination of features which is the best in distinguishing between true microRNA precursors and the background set. The score of each of the above features was divided into several score regions by a vector of thresholds from the minimal stringency to the maximal stringency (e.g. for hairpin length the vector is 55nts to 80nts in steps of 5nts). The different threshold vectors define a set of combinations of thresholds on all features. Each such combination may be viewed as a point on a directed graph where an edge exists from node A to node B if B is composed from the same set of thresholds as A except in one feature, and that feature is one step ahead (towards higher stringency) from that in A. The optimal predictor is the path on this graph from the least stringent node to the most stringent node, which optimizes the recall-precision balance, i.e. maximizes the percentage of known precursors and minimizes the percentage of background hairpins passing the thresholds defined by the nodes as measured on the whole path. The nodes in this path are given scores from 0.036 (least stringent node) to 1 (most stringent node) in steps of 0.036 (there are 28 nodes on

the optimal predictor). For each hairpin we find the highest stringency node on this optimal path which the hairpin passes its thresholds, and its score is the hairpin score, termed PalGrade. I.e. PalGrade=1 means that this hairpin scored in the top score region in all features, and PalGrade=0 means that the hairpin did not pass the minimal stringency criteria. Fig. 1B depicts the percentage of known precursors and of the background hairpins passing different PalGrade thresholds.

<u>MicroRNA sequence-directed cloning and sequencing</u>

Adaptors used in preparation of enriched microRNA libraries were: 5' adaptor (5' AACTGCAGAAAGGAGGAGCTCTAGrArTrA 3') and 3' adaptor ((5phos) rUrGrGAACAGATGAATTCTACC(3InvdT)). PCR amplification was performed with excess of the reverse primer (1:50 ratio), with primer: 5'TAATACGACTCACTATAGGTAGAATTCATCTGTTCCA3'. PCR conditions were: 4 minutes at 94°c followed by 7 cycles of 30 seconds at 93°c, 1 minute at 60°c, 30 seconds at 72 °c and then 30 cycles of 30 seconds at 93°c, 30 seconds at 60°c, and 30 seconds at 72 °c. The reaction was ended by incubation at 72°c for 10 minutes.
Hybridization was carried out in TEN buffer (10mM tris ph=8.0; 1mM EDTA; 100mM NaCl) for 1 hour at a temperature of Tm-10°c. µMACS Streptavidin Kit columns (130-074-101; Miltenyi Biotec, Gladbach, Germany) were processed according to the manufacturer instructions. PCR conditions for amplification of the recovered single-stranded cDNA library molecules were as described above. Vector used: #k1214, MBI Fermentas, Hanover, MD, USA.

<u>Estimating the number of microRNAs</u>

<u>I. Estimating the number of microRNAs using probabilistic arguments:</u>

Conserved hairpins: First, we define a highly conserved hairpin similar to a conserved hairpin (see Methods) except that the average score is at least 0.95 and the loop and flanking regions of the hairpin have a reduced score. Next, let S be a hairpin in a given score (PalGrade) group, C a conserved hairpin, HC a highly conserved hairpin, T a

microRNA gene, and F a hairpin which is not a microRNA gene. By the law of total probability P(HC|S,C) = P(HC|T,C)*P(T|S,C) + P(HC|F,C)*(1-P(T|S,C)). We find P(HC|S,C) empirically by computing the percent of highly conserved hairpins in the group of conserved hairpins in the specified PalGrade group, P(HC|T,C) = 0.97 by computing the percent of highly conserved hairpins in the known conserved microRNAs, and similarly P(HC|F,C) = 0.504 (using as F low scoring hairpins where true hairpins in the group leads to an underestimation of the final number). Given these empirical counts, we derive P(T|S,C), which is the expected percentage of true microRNAs in the PalGrade groups defined in Fig. 2 ("success rate"): 0, 0, 1.1, 5.21, and 18.75 respectively. To obtain the number of microRNAs in each group, P(T|S,C) is multiplied by its size in the entire genome: 46813, 23165, 8299, 1589 and 93 respectively. To obtain the final number of microRNAs, the total is divided by 0.86, the percentage of published microRNAs that are included in the initial candidate group of hairpins (see text). Summing over all groups gives the total estimated number of 222 conserved microRNAs in addition to the 220 known conserved microRNAs (total 442).

Non-conserved hairpins: First, we define a cluster of hairpins as a group (size>1) of hairpins all with PalGrade>0.3, with a maximal distance of 5000 nucleotides between 2 successive hairpins. A strong cluster is defined as a cluster in which at least one pair of hairpins exhibits a high sequence similarity (local alignment score of most similar arm is at least 12 using standard alignment with 1 for match, -1 for mismatch and -1 for gap opening and extension). Similar to the described above, we now use the equality P(T|MS,CL,NC) =[P(SCL|MS,CL,NC) - P(SCL|F,CL,NC)] / [P(SCL|T,CL,NC) - P(SCL|F,CL,NC)], where NC is not conserved, MS is PalGrade>0.3, CL is a hairpin in a cluster, and SCL is a hairpin in a strong cluster. We empirically find that P(SCL|MS,CL,NC) = 0.0375, P(SCL|F,CL,NC) = 0.0252, where again we use PalGrade<0.3 as the false group, and set P(SCL|T,CL,NC)=1 due to limited data, where this leads again to an underestimation of the final number since 1 is an overestimation of this probability. We then get P(T|MS,CL,NC)=0.0126, and since there are 19,948 non-conserved hairpins with PalGrade>0.3 in clusters, this gives 252 hairpins. Dividing by 0.75, which is the percent of published microRNAs that are in the initial candidate group

and have PalGrade>0.3, we derive the estimated number of 336 true microRNA genes in the group of non-conserved clustered microRNAs.

II. Estimating the number of conserved microRNAs based on sampling:

Conserved hairpins: Following are more details on Fig. 2. Samples were taken from highly conserved hairpins and then corrected for conserved hairpins in general to achieve better statistics. The 5 score groups are as depicted in Fig. 2, comprising of: 24811, 12587, 4416, 894 and 57 hairpins in the genome, respectively. Sample sizes used for validation via microarray followed by sequence-directed cloning were: 903, 489, 452, 216 and 24, respectively, from which 1, 4, 4, 18 and 5 cloned sequences were obtained, respectively. We deliberately underestimate it at 0, 1, 4, 18, and 5 successes. Four additional microRNAs not from the samples but having sequence similarity to the sequenced microRNAs were also sequenced but not considered for calculating success rates ('similar' microRNAs). The success rate in each group is the final number of sequenced microRNAs in the group (not including 'similar' microRNAs) divided by the sample size, i.e. the number of candidate microRNAs selected for microarray experiments (i.e. 0/903, 1/489, etc.), and dividing by 0.76 (see Figure 2), resulting in success rates of 0%, 0.3%, 1.2%, 11% and 27% respectively. The number of *conserved microRNAs* is obtained by multiplying the group size in the genome by its success rate, then dividing by the percent of published conserved microRNAs that are highly conserved (97%) to shift from highly conserved to conserved, and then by the percentage of published microRNAs that are in the initial candidate group (86%), yielding the final numbers of 0, 41, 62, 118, and 19 (total 240) conserved microRNAs (total 220+240=460).

Non-conserved hairpins: Non-conserved hairpins belonging to strong clusters with PalGrade groups of 3 to 5 as defined in fig. 2, were used, and comprised of: 553, 167 and 29 hairpins in the genome, respectively. Sample sizes for validation via microarray followed by sequence-directed cloning were: 22, 35, and 19 with sequence validation of: 1, 10, and 13 respectively. Additional 69 microRNAs from the clusters of the sequenced microRNAs, not appearing in the original samples, were sent to cloning

and sequencing, of which 29 were successfully sequenced ('adjacent microRNAs'), but not taken into consideration for calculating success rates. Success rates calculated as above (after division by 0.76 as above) were 6%, 38%, and 90%, respectively. The estimated number of true non-conserved clustered microRNAs is therefore obtained by multiplying the number of microRNAs in each group by its success rate, and dividing by the percentage of sequenced non-conserved microRNAs in clusters (96%) and by 86% as above, yielding the final estimated numbers of 43, 82, and 34 (total 159), respectively.

Reference

1. Bonnet, E. Wuyts J., Rouze, P. & Van de Peer, Y. Evidence that microRNA precursors unlike other non-coding RNAs have lower folding free energies than random sequences. *Bioinformatics,* **20(17)**, 2911-2917 (2004).

15. H. von Prahl and M. Alberico, *Isla de Gorgona* (Talleres Graficos Banco Popular, Bogota, 1986).
16. C. Birkeland, D. L. Meyer, J. P. Stames, C. L. Buford, *Smithson. Contrib. Knowl. Zool.* 176, 55 (1975).
17. H. G. Gierloff-Embden, *La Costa de El Salvador* (Direccion de Publicaciones, Ministerio de Educacion, San Salvador, 1976).
18. J. W. Durham, in *The Galapagos, Proceedings of the Symposia of the Galapagos International Scientific Project*, R. I. Bowman, Ed. (Univ. of California Press, Berkeley, 1966).
19. G. J. Bakus, *Atoll Res. Bull. 179* (1975), p. 1.
20. P. W. Glynn, *Environ. Conserv.* 10, 149 (1983); *ibid.* 11, 133 (1984).
21. Sixteen sites in the Gulf of Chiriqui were searched for *Millepora* spp. from 1984 to 1990. Two coral reef sites were surveyed intensively during 10 (Secas reef, Secas Islands) and 14 (Uva reef, Contreras Islands) different research cruises. These were sampled by (i) 30 permanent chain transects, each 10 m long [J. W. Porter, *Ecology* 53, 745 (1972)], with all corals counted in 300 transects, for a total search time of 44 diver hours; (ii) 12 permanent 1-m^2 and 20-m^2 quadrats with all corals mapped in a total of 620 m^2, 58 hours search time; and (iii) swimming surveys of forereef and reef base zones along each reef site, 40 hours. Forty hours of surveys at 14 other sites yields a total search effort for 1984 to 1990 of 204 hours.
22. R. H. Richmond, in *Global Ecological Consequences of the 1982–83 El Niño-Southern Oscillation*, P. W. Glynn, Ed. (Elsevier, Amsterdam, 1990).
23. P. W. Glynn, J. Cortes, H. M. Guzman, R. H. Richmond, *Proc. Sixth Intl. Coral Reef Symp.* 3, 237 (1988).
24. B. E. Brown and Suharsono, *Coral Reefs* 8, 163 (1990).
25. P. W. Glynn and L. D'Croz, *ibid.*, p. 181.
26. P. L. Jokiel, *Endeavour* 14, 66 (1990).
27. J. B. Lewis, *Coral Reefs* 8, 99 (1989).
28. G. C. Ray, in *Biodiversity*, E. O. Wilson, Ed. (National Academy Press, Washington, DC, 1988).
29. J. T. Carlton, G. J. Vermeij, D. R. Lindberg, D. A. Carlton, E. C. Dudley, *Biol. Bull.* 180, 72 (1991).
30. P. W. Glynn, in *Global Ecological Consequences of the 1982–83 El Niño-Southern Oscillation*, P. W. Glynn, Ed. (Elsevier, Amsterdam, 1990).
31. H. von Prahl and A. Mejia, *Rev. Biol. Trop.* 33, 39 (1985); H. von Prahl, personal communication.
32. H. M. Guzman, J. Cortes, P. W. Glynn, R. H. Richmond, *Mar. Ecol. Prog. Ser.* 60, 299 (1990).
33. G. Robinson, in *El Niño in the Galapagos Islands: The 1982–83 Event*, G. Robinson and E. M. del Pino, Eds. (Fundacion Charles Darwin para Las Islas Galapagos, Quito, Ecuador, 1985).
34. R. H. MacArthur, *Geographical Ecology* (Harper & Row, New York, 1972); E. G. Leigh Jr., *J. Theor. Biol.* 90, 213 (1981); J. M. Diamond, in *Extinctions*, M. H. Nitecki, Ed. (Univ. of Chicago Press, Chicago, 1984).
35. M. W. Colgan, in *Global Ecological Consequences of the 1982–83 El Niño-Southern Oscillation*, P. W. Glynn, Ed. (Elsevier, Amsterdam, 1990).
36. C. Emiliani, E. B. Kraus, E. M. Shoemaker, *Earth Planet. Sci. Lett.* 55, 317 (1981); E. G. Kauffman and C. C. Johnson, *Palaios* 3, 194 (1988).
37. S. M. Stanley, in *Extinctions*, M. H. Nitecki, Ed. (Univ. of Chicago Press, Chicago, 1984); J. A. Fagerstrom, *The Evolution of Reef Communities* (Wiley, New York, 1987).
38. Facilities and logistical support were provided by the Smithsonian Tropical Research Institute and the Centro de Ciencias del Mar y Limnologia, University of Panama. This paper benefited from discussions with D. P. de Sylva, C. M. Eakin, S. Snedaker, and H. R. Wanless. The authors acknowledge the support of NSF, grants OCE-8415615 and OCE-8716726 (P.W.G.), and the Research Opportunity Fund, Smithsonian Institution (W.H.d.W.). Permission to work in Panamanian waters was granted by the Direccion de Recursos Marinos, Ministerio de Comercio e Industrias, Republic of Panama. Contribution from the University of Miami, Rosenstiel School of Marine and Atmospheric Science.

6 March 1991; accepted 21 May 1991

Identification of the Envelope V3 Loop as the Primary Determinant of Cell Tropism in HIV-1

STEPHEN S. HWANG, TERENCE J. BOYLE, H. KIM LYERLY, BRYAN R. CULLEN*

Cells of the monocyte-macrophage lineage are targets for human immunodeficiency virus–1 (HIV-1) infection in vivo. However, many laboratory strains of HIV-1 that efficiently infect transformed T cell lines replicate poorly in macrophages. A 20–amino acid sequence from the macrophage-tropic BaL isolate of HIV-1 was sufficient to confer macrophage tropism on HTLV-IIIB, a T cell line–tropic isolate. This small sequence element is in the V3 loop, the envelope domain that is the principal neutralizing determinant of HIV-1. Thus, the V3 loop not only serves as a target of the host immune response but is also pivotal in determining HIV-1 tissue tropism.

Although the $CD4^+$ lymphocyte is the major target for HIV-1 replication in the peripheral blood compartment, cells of the monocyte-macrophage lineage represent the predominant HIV-1–infected cell type in most tissues, including the central nervous system (*1–3*). An HIV-1 infection of macrophages, although less cytopathic than an infection of T cells, compromises macrophage function and may underlie many of the pathogenic effects of HIV-1 infection in humans (*1–3*). Despite the importance of macrophages as a primary target for HIV-1 infection in vivo, many laboratory isolates of HIV-1 are unable to replicate in these cells (*3–8*).

The HIV-1 isolates can be divided into two major subgroups on the basis of their cellular host range in vitro (*2–9*). Macrophage–T cell (MT)–tropic isolates efficiently infect both macrophages and $CD4^+$ peripheral blood lymphocytes (PBLs) but are unable to replicate efficiently in many transformed cell lines of either T cell or monocytic origin. A second class of viruses, termed T cell (T)–tropic, replicates efficiently in both PBLs and transformed T cell lines but poorly in macrophages (*2–9*). The tropism of HIV-1 is determined early in the viral replication cycle, between binding of the virus to the cell surface and initiation of viral reverse transcription (*6, 7*) and is independent of the viral long terminal repeat (*8*) but dependent on sequences in the viral gp120 envelope protein (*5, 6*).

The HTLV-IIIB isolate of HIV-1 is T-tropic, whereas the BaL strain is MT-tropic (*3, 8*). To determine which sequences were important for in vitro tropism, we cloned and sequenced the envelope gene and flanking viral sequences of BaL with DNA derived from BaL-infected macrophages (*10*). We then constructed a series of chimeric HIV-1 proviruses by substitution of BaL-derived sequences into an HTLV-IIIB proviral clone (Fig. 1). These chimeric viruses were then tested for tropism by analysis of their replication competence in PBLs (*11*), primary monocyte-derived macrophages (*12*), and the transformed $CD4^+$ T cell lines H9 and CEM (Table 1).

Both the parental HTLV-IIIB clone (pIIIB) and the provirus with the complete BaL *env* gene (pBaL) replicated equivalently in primary PBLs, as we determined by measuring secreted p24 Gag protein (Table 1) or supernatant reverse transcriptase activity (*13*). As predicted (*3, 8*), the pIIIB provirus also replicated efficiently in both the H9 and CEM cell lines but not in macrophages. In contrast, the pBaL provirus replicated efficiently in macrophages but not in H9 or CEM cells (Table 1). The various chimeric proviral constructs, like the parental pIIIB and pBaL clones, all displayed comparable replication competence in PBLs.

In addition, all chimeric clones were either fully T-tropic or fully MT-tropic; that is, no intermediate or dual tropism was detected. As previously reported (*5, 6*), tropism was determined entirely by sequences located within gp120. A 20–amino acid

S. S. Hwang, Department of Microbiology and Immunology, Duke University Medical Center, Durham, NC 27710.
T. J. Boyle and H. K. Lyerly, Department of Surgery, Duke University Medical Center, Durham, NC 27710.
B. R. Cullen, Howard Hughes Medical Institute and Department of Microbiology and Immunology, Duke University Medical Center, Durham, NC 27710.

*To whom correspondence should be addressed.

BaL-specific sequence that introduces only 11 amino acid changes into the pIIIB envelope protein was both necessary and sufficient to confer a fully MT-tropic phenotype. An HTLV-IIIB–derived provirus bearing this minimal substitution, termed IIIB/V3-BaL, was observed to replicate in macrophages at least as well as the pBaL proviral clone, yet could no longer infect T cell lines (Table 1). The introduced 20–amino acid BaL-specific sequence coincides with the core of the V3 loop of the HIV-1 envelope protein (Table 2), a discrete ~35–amino acid protein domain that forms the principal neutralizing determinant of HIV-1 (*14–16*). Our data demonstrate that the V3 loop is also a major determinant of HIV-1 cell tropism.

A 159–amino acid envelope sequence has been identified as a critical determinant of HIV-1 tropism (*5, 6*). This longer sequence extends 101 amino acids NH_2-terminal and 38 amino acids COOH-terminal to the 20–amino acid V3 loop sequence defined here, suggesting that the V3 loop may also be a major determinant of tropism in other viral isolates. These experimentally defined MT-tropic V3 loop sequences are all similar to each other and to the statistically prevalent or "consensus" V3 loop sequence defined by examination of 245 distinct isolates of HIV-1 (*16*) (Table 2). In contrast, T-tropic isolates appear to be characterized by statistically rare V3 loop sequences that are dissimilar to the consensus. It therefore appears possible that envelope V3 loop sequences may help predict the cell tropism of primary HIV-1 isolates.

MT-tropic isolates are the predominant HIV-1 class detected early after infection of humans, whereas T-tropic isolates become more prevalent as disease progresses (*9*). This progression may be involved in HIV-1 pathogenesis (*9*). Similarly, the V3 loop sequence is also subject to rapid change both within and between different human hosts (*15, 16*). This evolution has been ascribed to the immunological selection of virus-bearing variations in the V3 loop sequence that permit escape from V3 loop–specific neutralizing antibodies (*15*). However, our data suggest that V3 loop evolution may also reflect the selection of T-tropic variants with nonconsensus V3 loop sequences during the later stages of HIV-1–induced disease. Therefore, vaccines intended to protect against challenge by the MT-tropic isolates prevalent during the early stages of HIV-1 infection should perhaps be designed to elicit an immune response specific for V3 loop sequences similar to the consensus.

The mechanism by which the V3 loop influences HIV-1 cell tropism remains unclear. The V3 loop probably does not interact with a primary cell surface receptor that is distinct from CD4. CD4 binding is critical for infection of both macrophages and T cells (*17*), and a defective or occluded V3 loop does not affect the ability of gp120 to bind cell surface CD4 (*14*). Thus, the V3 loop is likely to be involved in a step immediately subsequent to the initial gp120-CD4 binding event, which results in the activation of the fusogenic potential embodied in the hydrophobic NH_2-terminus of the gp41 component of the envelope (*18*).

An activation step would preclude the premature fusion of gp41 with cell membranes encountered during the intracellular posttranslational processing of envelope and would also prevent the random fusion of virions with $CD4^-$ cells unable to support viral replication, such as reticulocytes (*18, 19*). In some enveloped animal viruses, activation of fusion occurs subsequent to a low pH-induced conformational change in the viral membrane protein after endocytosis of the virion (*19, 20*). However, fusion of HIV-1 with $CD4^+$ cells requires neither exposure to low pH nor internalization of the CD4 receptor (*20, 21*). Alternatively, envelope fusion in HIV-1 might require a specific proteolytic activation step, as has been suggested for a number of enveloped viruses, including the retrovirus murine leu-

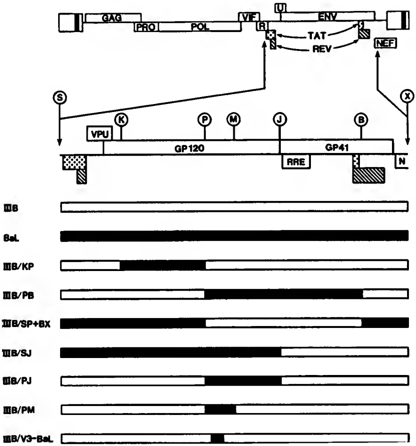


Fig. 1. Structure of chimeric HIV-1 proviral clones. The HXB-3 proviral clone is full length and replication-competent and is derived from the HTLV-IIIB isolate of HIV-1 (*26*). The HXB-3 provirus present in the pIIIB plasmid is similar to the published clone except that the single base pair frame-shift mutation present in the *vpr* gene of HXB-3 has been repaired (*27*). In the majority of cases, BaL sequences (in black) were substituted into the pIIIB clone (in white) after digestion at the indicated restriction enzyme sites. However, some proviral chimeras were constructed with polymerase chain reaction primers specific for the gp120-gp41 junction or for the V3 loop region, in combination with flanking primers homologous to sequences adjacent to the unique viral Kpn I and Bam HI sites (*10*). In the case of pIIIB/V3-BaL, the primers used were 5′-GGGTCCTATATGTATACTTTTTCTTG-TATTGTTGTTGGG-3′ and 5′-AAAAAGTATACATATAGGACCCGGGAGAGCATTATATACA-ACAGGAGAAATAATAGGAGATATAAG-3′. These primers permitted the construction of a substitution mutant, pIIIB/V3-BaL, that is identical to the pIIIB parent except for a region of 20 amino acids in the core of the V3 loop sequence (Table 2). The origin of the various chimeric proviral clones is indicated in the plasmid name given at left (for example, IIIB/KP contains a BaL-derived Kpn I to Pvu II fragment) and was confirmed by dideoxynucleotide sequence analysis. K, Kpn I; P, Pvu II; M, Mst II; J, gp120-gp41 junction; B, Bam HI; RRE, Rev Response Element; R, Vpr; U, Vpu; N, Nef; S, Sal I; X, Xho I.

Table 1. Cell tropism of chimeric HIV-1 proviral clones as measured by p24 Gag expression in picograms per milliliter. To assess tropism, we transfected HIV-1 proviral clones (*25*) into cultures of the monkey cell line COS that were in 35-mm plates. At 3 days after transfection, the supernatant media was replaced by 2 ml of RPMI medium with PHA-stimulated PBLs (2×10^6) (*11*). Activated PBLs were cultivated with the transfected COS cells for 3 days, aspirated, washed, and maintained in expanded culture for four more days. On day 7 (d7) after infection, supernates (10 ml) were removed from the PBLs and filtered through a 0.45 μM filter, and p24 was assayed by enzyme-linked immunosorbent assay (Du Pont Biotechnology Systems). These values are given in the column marked "PBL." The p24 levels were then standardized by dilution to ~200 pg/ml (experiment 1) or ~300 pg/ml (experiment 2), and we used 500 μl of each virus supernatant to infect adherent monocyte-derived macrophages (MØ) (*12*), 5×10^5 H9 cells, or 5×10^5 CEM cells. Supernatant media were replaced twice per week and monitored for p24 expression levels. The data given are for 14 days (d14) after infection, but similar results were also routinely obtained at days 7 and 21. No p24 expression was detected in negative control cultures.

Chimera	Experiment 1				Experiment 2		
	PBL d7	MØ d14	H9 d14	CEM d14	PBL d7	MØ d14	H9 d14
IIIB	306	<3	1368	474	572	11	138
BaL	341	627	<3	<3	296	298	5
IIIB/KP	368	<3	927	442	696	5	298
IIIB/PB	272	1161	<3	<3	322	301	5
IIIB/SP+BX	356	<3	505	629	585	3	344
IIIB/SJ	248	694	<3	<3	564	259	<3
IIIB/PJ	347	1250	<3	<3	340	252	3
IIIB/PM	366	1426	<3	<3	448	616	<3
IIIB/V3-BaL	291	2172	<3	<3	364	709	<3

Table 2. Comparison of V3 loop sequences of HIV-1 isolates of known tropism. The ~35–amino acid V3 loop is a discrete envelope protein domain defined by two invariant, disulfide-bonded cysteine residues (*15*). A statistically prevalent or consensus V3 loop sequence has been defined (*16*) and this is identical to the V3 loop of the MT-tropic isolate JR-FL (*6*). Modification of the HTLV-IIIB V3 loop sequence to match that of the BaL isolate, by exchange of the boxed amino acids, is sufficient to confer an MT-tropic phenotype. This mutation introduces 11 amino acid changes into the envelope protein of HTLV-IIIB. Abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. Each dash indicates identity with the BaL sequence; each dot, a deletion.

Clone	Sequence	Tropism
SF-2	---------- --Y-••-----FH---R----- -K---	T
NL4-3	---------- --R-QR-----FV-I-K•--NM -----	T
HTLV-IIIB	---------- K-R-QR-----FV-I-K•--NM -----	T
BaL	CTRPNNNTRK SIHI••GPGRALYTTGEIIGDI RQAHC	MT
JR-FL (consensus)	---------- ----•------F---------- -----	MT
SF162	---------- --T-••-----F-A--D----- -----	MT

kemia virus (*20, 22, 23*). HIV-1 tropism would then reflect the availability of cell surface or lysosomal proteases capable of cleaving specific sites within different viral V3 loops. Evidence suggesting that the V3 loop is a target for sequence-specific proteases has been presented (*24*), and the availability of appropriate cellular proteases is known to affect tissue tropism in some other viral species (*23*). If the sequence-specific cleavage of V3 is indeed critical for HIV-1 infection, then this site might well provide a novel and attractive target for chemotherapeutic intervention in HIV-1–induced disease.

REFERENCES AND NOTES

1. M. H. Stoler, T. A. Eskin, S. Benn, R. Angerer, L. M. Angerer, *J. Am. Med. Assoc.* **256**, 2360 (1986); S. Koenig *et al.*, *Science* **233**, 1089 (1986); Y. Koyanagi *et al.*, *ibid.* **236**, 819 (1987); S. M. Schnittman *et al.*, *ibid.* **245**, 305 (1989); C. D. Pauza and T. M. Price, *J. Cell Biol.* **107**, 959 (1989); C. Wiley, R. Schrier, J. Nelson, P. Lampert, M. Oldstone, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7089 (1986); G. C. Baldwin *et al.*, *ibid.* **87**, 3933 (1990); B. J. Potts, W. Maury, M. A. Martin, *Virology* **175**, 465 (1990).
2. H. E. Gendelman *et al.*, *AIDS* **3**, 475 (1989); M. S. Meltzer, D. R. Skillman, P. J. Gomatos, D. C. Kalter, H. E. Gendelman, *Annu. Rev. Immunol.* **8**, 169 (1990).
3. S. Gartner *et al.*, *Science* **233**, 215 (1986).
4. L. A. Evans, T. M. McHugh, D. P. Stites, J. A. Levy, *J. Immunol.* **138**, 3415 (1987); R. Collman, *J. Exp. Med.* **170**, 1149 (1989); C. Cheng-Mayer, M. Quiroga, J. W. Tung, D. Dina, J. A. Levy, *J. Virol.* **64**, 4390 (1990).
5. T. Shioda, J. A. Levy, C. Cheng-Meyer, *Nature* **349**, 167 (1991).
6. W. A. O'Brien *et al.*, *ibid.* **348**, 69 (1990).
7. S. Kim, K. Ikeuchi, J. Groopman, D. Baltimore, *J. Virol.* **64**, 5600 (1990).
8. R. J. Pomerantz, M. B. Feinberg, R. Andino, B. Baltimore, *ibid.* **65**, 1041 (1991).
9. H. Schuitemaker *et al.*, *ibid.*, p. 356; C. Cheng-Mayer, D. Seto, M. Tateno, J. A. Levy, *Science* **240**, 80 (1988); E. M. Fenyo *et al.*, *J. Virol.* **62**, 4414 (1988); M. Tersmette *et al.*, *ibid.* **63**, 2118 (1989).
10. The envelope gene of BaL was cloned in segments by the polymerase chain reaction [K. B. Mullis and F. A. Faloona, *Methods Enzymol.* **155**, 335 (1987)] as with a series of 20 nucleotide primers homologous to conserved HIV-1 sequences adjacent to the proviral Sal 1, Kpn 1, Pvu II, Bam HI, and Xho I restriction enzyme sites indicated in Fig. 1. The nucleotide sequence of this BaL-specific segment can be obtained from GenBank under accession number M63929.
11. Peripheral mononuclear cells (PMCs) were separated from normal donor peripheral blood by a Ficoll-Hypaque (Sigma) gradient [A. Boyum, *Scand. J. Clin. Lab. Invest. Suppl.* **97**, 1 (1968)]. PBLs were separated from plastic adherent PMCs by cultivation in RPMI 1640 medium containing 20% fetal calf serum, interleukin-2 (20 units/ml), and phytohemagglutinin (PHA) (5 μg/ml) for 72 hours [P. C. Nowell, *Cancer Res.* **20**, 562 (1960)]. PBLs were ~50% $CD4^+$ by flow cytofluorimetry.
12. After isolation of PMCs by Ficoll-Hypaque gradient, an ~80% monocyte subpopulation (determined by flow cytofluorimetry) was isolated by discontinuous Percoll (Pharmacia-LKB) gradient centrifugation [A. J. Ulmer and H. D. Flad, *J. Immunol. Methods* **30**, 1 (1979)]. Cells (10^6) were plated into 48-well plates with RPMI 1640 medium containing 20% Human AB serum. After allowing the cells to adhere to plastic (1 hour) and washing them [H. Pertoft *et al.*, *ibid.* **33**, 221 (1980)], we obtained a 99% pure population of monocytes, as determined by nonspecific esterase staining. We maintained cells for 5 days in culture to permit differentiation to macrophages before we initiated infection.
13. T. J. Boyle, unpublished observations.
14. M. A. Skinner *et al.*, *J. Virol.* **62**, 4195 (1988); E. O. Freed, D. J. Myers, R. Risser, *ibid.* **65**, 190 (1991).
15. T. J. Palker *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1932 (1988); J. R. Rusche *et al.*, *ibid.*, p. 3198; J. Goudsmit *et al.*, *ibid.*, p. 4478; T. F. W. Wolfs *et al.*, *ibid.* **87**, 9938 (1990); K. Javaherian *et al.*, *Science* **250**, 1590 (1990).
16. G. J. LaRosa *et al.*, *Science* **249**, 932 (1990).
17. D. H. Smith *et al.*, *ibid.* **238**, 1704 (1987); R. A. Fisher *et al.*, *Nature* **331**, 76 (1988); K. C. Deen *et al.*, *ibid.*, p. 82; P. R. Clapham *et al.*, *ibid.* **337**, 368 (1989); R. Collman *et al.*, *J. Virol.* **64**, 4468 (1990).
18. J. M. McCune *et al.*, *Cell* **53**, 55 (1988); D. Camerini and B. Seed, *ibid.* **60**, 747 (1990); R. L. Willey, E. K. Ross, A. J. Buckler-White, T. S. Theodore, M. A. Martin, *J. Virol.* **63**, 3595 (1989); J. P. Moore, J. A. McKeating, W. A. Norton, Q. J. Sattentau, *ibid.* **65**, 1133 (1991).
19. J. White, M. Kielian, A. Helenius, *Q. Rev. Biophys.* **16**, 151 (1983); P. W. Choppin and A. Scheid, *Rev. Infect. Dis.* **2**, 40 (1980).
20. M. O. McClure, M. A. Sommerfelt, M. Marsh, R. A. Weiss, *J. Gen. Virol.* **71**, 767 (1990).
21. B. S. Stein *et al.*, *Cell* **49**, 659 (1987); P. J. Maddon *et al.*, *ibid.* **54**, 865 (1988); M. O. McClure, M. Marsh, R. A. Weiss, *EMBO J.* **7**, 513 (1988).
22. K. Otsuki and M. Tsubokura, *Arch. Virol.* **70**, 315 (1981); H. Yoshikura and S. Tejima, *Virology* **113**, 503 (1981); G. Appleyard and M. Tisdale, *J. Gen. Virol.* **66**, 363 (1985); K. B. Andersen, *ibid.* **68**, 2193 (1987); ______ and H. Skov, *ibid.* **70**, 1921 (1989).
23. A. Scheid and P. W. Choppin, *Virology* **69**, 265

(1976); T. E. Toth, *Am. J. Vet. Res.* **43**, 967 (1982).
24. T. Hattori, A. Koito, K. Takatsuki, H. Kido, N. Katunuma, *FEBS Lett.* **248**, 48 (1989); P. E. Stephens, G. J. Clements, G. T. Yarranton, *Nature* **343**, 219 (1990); G. J. Clements *et al.*, *AIDS Res. Hum. Retroviruses* **7**, 3 (1991).
25. B. R. Cullen, *Methods Enzymol.* **152**, 684 (1987).
26. G. M. Shaw *et al.*, *Science* **226**, 1165 (1984); R. Crowl *et al.*, *Cell* **41**, 979 (1985); M. H. Malim, S. Bohnlein, J. Hauber, B. R. Cullen, *ibid.* **58**, 205 (1989).
27. E. D. Garrett, L. S. Tiley, B. R. Cullen, *J. Virol.* **65**, 1653 (1991).
28. We thank M. Malim, L. Tiley, and A. Langlois for helpful discussions and S. Goodwin for secretarial assistance.

4 March 1991; accepted 22 May 1991

The *ras* Oncoprotein and M-Phase Activity

IRA DAAR, ANGEL R. NEBREDA, NELSON YEW, PHILIP SASS, RICHARD PAULES,* EUGENIO SANTOS, MICHAEL WIGLER, GEORGE F. VANDE WOUDE†

The endogenous *mos* proto-oncogene product (Mos) is required for meiotic maturation. In *Xenopus* oocytes, the *ras* oncogene product (Ras) can induce meiotic maturation and high levels of M-phase–promoting factor (MPF) independent of endogenous Mos, indicating that a parallel pathway to metaphase exists. In addition, Ras, like Mos and cytostatic factor, can arrest *Xenopus* embryonic cell cleavage in mitosis and maintain high levels of MPF. Thus, in the *Xenopus* oocyte and embryo systems Ras functions in the M phase of the cell cycle. The embryonic cleavage arrest assay is a rapid and sensitive test for Ras function.

IN THE *XENOPUS LAEVIS* SYSTEM, FULLY grown oocytes are arrested in prophase of the first meiotic division. Progesterone releases this arrest, resulting in activation of MPF, germinal vesicle breakdown (GVBD), completion of meiosis I, and production of an unfertilized egg arrested at metaphase II of meiosis (*1*). MPF is composed of the *Xenopus* homolog of the cell cycle regulator $p34^{cdc2}$ and cyclin (*2*) and is present at high levels in unfertilized eggs (*1*). Cytostatic factor (CSF) is also found in unfertilized eggs and is believed to be responsible for the arrest of maturation at metaphase II of meiosis (*1, 3*). Mos has been shown to be an active component of CSF (*4*), and introduction of CSF or Mos into the blastomeres of rapidly cleaving embryos arrests cleavage at metaphase of mitosis (*1, 3, 4*). This arrest by CSF or Mos, at a major cell cycle control point (*5*), results from the stabilization of high levels of MPF (*3, 4, 6, 7*).

The unrestricted proliferation of cells transformed by oncogenes provides a strong argument that proto-oncogenes normally function in the regulation of the cell cycle (*8*). Research emphasis has been directed toward understanding how oncogenes alter the regulation of signal transduction events in the G_0 to G_1 phase of the cell cycle (*9*). The discovery that Mos functions during M phase (*4, 10*) led us to propose that the transforming activity of the Mos in somatic cells is due to the expression of its M-phase activity during interphase (*4, 10, 11*). A similar hypothesis has been presented for the *src*-transforming activity (*12*), and this may be a more general mechanism for how certain oncogenes induce morphological transformation (*4, 10, 11*).

Ras, the transforming guanosine triphosphate (GTP)–binding protein (*13*), and Mos induce progesterone-independent meiotic maturation in *Xenopus* oocytes (*11, 14–17*) (Table 1). We tested Ras in this assay by injecting either $\text{Ras}^{\text{Lys12}}$ or $\text{H-Ras}^{\text{Val12}}$ RNA. Injected oocytes were subsequently examined for GVBD and MPF activity (*18*). Cytosolic extracts prepared from oocytes induced to mature with these products were positive for MPF, indicating that the oocytes were arrested in metaphase (Table 1). In addition, these analyses confirm that Ras (*19*), like Mos, can sustain high levels of MPF after GVBD (Table 1).

In fully grown *Xenopus* oocytes, antisense oligodeoxyribonucleotides destabilize the *mos* maternal mRNA and block progesterone-induced meiotic maturation (*10, 15*). To test whether Ras could induce meiotic maturation in the absence of progesterone and endogenous *mos* mRNA, we injected *mos*-specific antisense or sense oligodeoxyribonucleotides (*10*) into oocytes 3.5 to 4 hours before injecting the test material and subsequently examined them for GVBD and MPF activity (Table 1). GVBD occurred frequently in Mos-negative oocytes injected with Ras (60%), and extracts prepared from oocytes that displayed GVBD were positive for MPF activity (Table 1). Barrett and co-workers have shown that Mos depletion inhibits Ras-induced maturation (*15*). Allende and co-workers reported that Ras can induce GVBD in cycloheximide-treated oocytes (*16*), and Barrett and co-workers also observed this on occasion (*15*). These latter results are more consistent with our data because Mos is not synthesized in oocytes in the presence of cycloheximide (*11, 20*). Moreover, Ras-induced oocyte maturation appears to be Mos-dependent in less mature Dumont stage V (*21*) oocytes, but not in fully grown stage VI oocytes (*22*), presumably because of metabolic changes during oogenesis.

Because Ras induces meiotic maturation and high levels of MPF in oocytes, we tested whether it influences M-phase events in cleaving embryos, where the cell cycle consists essentially of S and M phases. Ras efficiently arrested embryonic cleavage when one blastomere of each two-cell embryo was injected with either oncogenic Ras protein or RNA (Figs. 1 and 2). This cleavage arrest mimics the arrest caused by CSF or Mos (*4*). Moreover, as little as 1 to 2 ng of Ras could induce the cleavage arrest, which was observable within a few hours (Fig. 2).

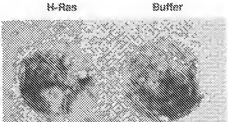


Fig. 1. Morphology of embryos injected with $\text{H-Ras}^{\text{Val12}}$ RNA. Animal-pole view of embryos injected with either capped $\text{H-Ras}^{\text{Val12}}$ RNA (*18*) or buffer. The RNA or buffer was microinjected into one blastomere (bottom half) of a two-cell embryo and examined several hours later.

Although transforming Ras induced the cessation of embryonic cleavage, both normal and nontransforming mutant forms of Ras had no observable effect on cleavage, even when introduced at concentrations approximately ten times the minimum effective dose for the transforming Ras. Thus, 15 ng of either normal Ras or $\text{Ras}^{\text{Lys12Ser186}}$, a protein that cannot associate with the plas-

I. Daar, N. Yew, R. Paules, G. F. Vande Woude, ABL–Basic Research Program, National Cancer Institute–Frederick Cancer Research and Development Center, Frederick, MD 21701.
A. R. Nebreda and E. Santos, National Institute of Allergy and Infectious Diseases, Laboratory of Molecular Microbiology, Bethesda, MD 20892.
P. Sass, Lederle Laboratory, Pearl River, NY 10965.
M. Wigler, Mammalian Cell Genetics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724.

*Present address: Mammalian Molecular Genetics of Carcinogenesis Group, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709.
†To whom correspondence should be addressed.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 280, No. 30, Issue of July 29, pp. 27595–27603, 2005
Printed in U.S.A.

Efficient Processing of Primary microRNA Hairpins by Drosha Requires Flanking Nonstructured RNA Sequences*

Received for publication, April 29, 2005, and in revised form, May 31, 2005
Published, JBC Papers in Press, June 1, 2005, DOI 10.1074/jbc.M504714200

Yan Zeng‡ and Bryan R. Cullen§

From the Center for Virology and Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina 27710

Drosha is a member of the ribonuclease (RNase) III family that selectively processes RNAs with prominent double-stranded features. Drosha plays a key role in the generation of precursor microRNAs from primary microRNA (pri-miRNA) transcripts in animal cells, yet how Drosha recognizes its RNA substrates remains incompletely understood. Previous studies have indicated that, within the context of a larger pri-miRNA, an ~80-nucleotide-long RNA hairpin structure is necessary for processing by Drosha. Here, by performing *in vitro* Drosha processing reactions with RNA substrates of various sizes and structures, we show that Drosha function also requires single-stranded RNA extensions located outside the pri-miRNA hairpin. The sequence of these RNA extensions was largely unimportant, but a strong secondary structure within the extension or a blunt-ended pri-miRNA hairpin blocked Drosha cleavage. The requirement for single-stranded extensions on the pri-miRNA hairpin substrate for Drosha processing is currently unique among the RNase III enzymes.

Ribonuclease (RNase) III family enzymes are expressed in both prokaryotes and eukaryotes and are involved in the processing, maturation, and degradation of a wide variety of RNAs, including ribosomal RNAs, transfer RNAs, small nuclear RNAs, small nucleolar RNAs, microRNAs (miRNAs),[1] and small interfering RNAs (1, 2). These RNases generate RNA products that feature an imperfect or perfect duplex with an ~2-nucleotide (nt) 3′ overhang at the site of cleavage. This characteristic staggered 3′ end structure results from the independent cleavage of the two RNA strands by the two catalytic sites located within a single double-stranded RNA (dsRNA) processing center formed by two RNase III domains. These RNase III domains may derive from two proteins, as seen with bacterial RNase III, or from a single protein, as seen with Dicer and Drosha (3, 4).

The RNase III family can be divided into four subclasses (1). Class 1 consists of bacterial enzymes with a minimal RNase III domain and a single dsRNA binding domain (dsRBD). Class 2 consists of fungal enzymes such as Rnt1p in *Saccharomyces cerevisiae* and PacI in *Schizosaccharomyces pombe*, which contain an extra N-terminal region with no recognizable motifs. Class 3 consists of the Drosha orthologs found in animals. These proteins have two RNase III domains and one dsRBD in the C-terminal half and a proline-rich domain and an arginine-rich (R-rich) domain in the N-terminal half of the protein. Class 4 RNase III enzymes consist of the Dicer homologs expressed in *S. pombe*, plants, and animals. Their C-terminal half appears similar to Drosha, but the N-terminal half features different domain structures.

How these RNase III enzymes select and cleave their RNA substrates has been the subject of several studies. *Escherichia coli* RNase III, a class 1 enzyme, targets a broad spectrum of RNAs, apparently regulated only by RNA antideterminants, *i.e.* certain sequences arranged in base pairs at defined positions along a helical substrate are disfavored (5). Human Dicer, a class 4 enzyme, preferentially recognizes a 2-nt 3′ overhang on a dsRNA and then cuts ~20 nt away to generate a short RNA duplex (3, 6, 7). Rnt1p, a class 2 enzyme, shows the highest specificity, as it selects for a NGNN (N, any nucleotide) tetraloop and cleaves 14–16 bp into the stem of the flanking RNA hairpin (8, 9). Another class 2 enzyme, PacI, however, does not have such stringent requirements (10). Notably, although the enzymes mentioned above show clearly distinct substrate specificities, they are all capable of processing a blunt-ended dsRNA substrate effectively *in vitro*.

The class 3 RNase III Drosha forms a complex with a protein partner, termed DGCR8 in humans and Pasha in flies and worms, that catalyzes the cleavage of long primary miRNA transcripts (pri-miRNAs) to produce the ~60-nt hairpin RNAs termed precursor miRNAs (pre-miRNAs) (4, 11–13). Pre-miRNAs are further processed by Dicer to yield mature, ~22-nt-long miRNAs. miRNAs are abundant, endogenous, noncoding RNAs that post-transcriptionally regulate gene expression in multicellular organisms (14). Because Drosha produces pre-miRNAs that then serve as substrates for Dicer and because Dicer primarily uses the terminal structure of the pre-miRNA hairpin created by Drosha cleavage to determine where it will subsequently cut, pri-miRNA cleavage by Drosha imparts much of the specificity of miRNA processing in animal cells.

It has been demonstrated that for a pri-miRNA to be efficiently processed by Drosha the targeted hairpin must consist of a large terminal loop of ≥10 nt and a stem region somewhat longer than the one present in the final pre-miRNA (11, 15, 16). In all of the previously reported experiments that analyzed Drosha activities *in vitro* and miRNA expression in transfected cells, the miRNA-containing hairpin was always embedded within a longer transcript and thus surrounded by extra RNA sequences derived either from its endogenous flanking genomic

* This work was supported by National Institutes of Health Grant GM071408 (to B. R. C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Present address: Dept. of Radiation Oncology, Virginia Commonwealth University, 401 College St., Richmond, VA 23298.

§ To whom correspondence should be addressed: Dept. of Molecular Genetics and Microbiology, Duke University Medical Center, Box 3025, Durham, NC 27710. Tel.: 919-684-3369; Fax: 919-681-8979; E-mail: culle002@mc.duke.edu.

[1] The abbreviations used are: miRNA, microRNA; nt, nucleotide(s); dsRNA, double-stranded RNA; RBD, RNA binding domain; dsRBD, double-stranded RNA binding domain; R-rich, arginine-rich; pri-miRNA, primary microRNA transcript; pre-miRNA, precursor microRNA; ssRNA, single-stranded RNA; GST, glutathione *S*-transferase.

sequence or from the expression vector used (4, 11, 15–18). One of these studies shows that, when transcribed from an RNA polymerase III promoter in transfected cells, at least 40 nt of additional sequence on each side of a pre-miRNA structure is required for efficient miRNA production (18). Some of the flanking nucleotides likely formed the short stem extension beyond the pre-miRNA stem that is known to be essential for pri-miRNA processing (11, 15, 16), but what role the rest of these extra sequences play remains unclear. Do they provide specific RNA sequences or structures that enhance processing, or does this simply represent a requirement for flanking single-stranded RNA (ssRNA)? Furthermore, little is known about how Drosha interacts with the various structural elements required for efficient pri-miRNA processing. To address these questions, we have analyzed how various flanking RNA sequences affect pri-miRNA hairpin recognition and cleavage by Drosha.

MATERIALS AND METHODS

Plasmid Construction—pGEX-4T-1-Drosha RBD encodes part of the Drosha protein extending from leucine 1254 to its C terminus and was made by amplifying the relevant DNA fragment with primers 5′-CTG-GAATTCATGTTGAATCAGGATTGGAAT-3′ and 5′-GCGCTCGAGT-TATTTCTTGATGTCTTCAGT-3′, digesting the PCR product with EcoRI and XhoI, and inserting it into the EcoRI and XhoI sites of pGEX-4T-1. pGEX-4T-1-Drosha R-rich encodes the sequence from proline 216 to leucine 333 of Drosha and was similarly made by PCR subcloning using primers 5′-GCGAATTCCCCAGTGAGAGAAGGTC-C-3′ and 5′-GCCTCGAGCTAATTCTGGTGTGCATCC-3′. pGEX-4T-1-DGCR8 2XRBD encodes a DGCR8 protein fragment extending from glutamic acid 502 to the C terminus and was made by amplifying a DNA fragment from a FLAG-DGCR8 expression plasmid (a gift of Dr. R. Shiekhattar) with primers 5′-GCGAATTCGAGTTTGTTATTAACCCC-AAC-3′ and 5′-GAGGTCGACGTTAACTCACACGTCCACGGTGCACA-3′, digesting the DNA with EcoRI and SalI, and cloning into the EcoRI and XhoI sites of pGEX-4T-1.

Preparation of Recombinant Drosha—Human 293T cells were transfected with pCK-Drosha-FLAG, which expresses C-terminally FLAG-tagged human Drosha (11). Two days later, cell extracts were prepared in lysis buffer (20 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, and 0.4% Nonidet P-40) and mixed with anti-FLAG-agarose beads (Sigma) at 4 °C for ~1 h. Beads were washed four times with the same lysis buffer and once with reaction buffer (20 mM HEPES-KOH, pH 7.6, 100 mM KCl, 0.2 mM EDTA, and 5% glycerol). Proteins were then eluted by incubating the beads with reaction buffer containing 150 ng/μl 3× FLAG peptide (Sigma) at 4 °C for ~30 min. Aliquoted supernatants were stored at −80 °C.

Enzymatic Assays—Drosha processing experiments were performed in the reaction buffer mentioned above supplemented with 2 mM dithiothreitol, 7 mM $MgCl_2$, and ~0.5 unit/μl RNasin (Promega). Purified Drosha was mixed with ~10^4 cpm (~0.1 ng) of a ^{32}P-labeled RNA substrate and incubated at 37 °C for 60–90 min. Reactions were terminated by adding an equal volume of 2× loading buffer (98% formamide, 20 mM EDTA, and 0.1% bromphenol blue), heated at 95 °C for ~10 min, resolved on a 10% denaturing gel, and analyzed by autoradiography and/or on a PhosphorImager (Amersham Biosciences). 5′-^{32}P-Labeled φX174 HinfI DNA markers (Promega) were used as size standards. All RNA substrates were tested at least twice.

DNA templates for RNA probes were prepared by PCR using T7 promoter-added primers. RNA probes were transcribed *in vitro* (Promega) in the presence of [α-^{32}P]CTP. For some of the probes, full-length RNAs were gel-isolated prior to use.

To prepare the circular primary miR-31 substrate, the linear, labeled transcript was first synthesized as described above, dephosphorylated by alkaline phosphatase (Roche Applied Science), purified, phosphorylated by T4 polynucleotide kinase (New England Biolabs) in the presence of ATP, and then treated with T4 RNA ligase (New England Biolabs). Products were separated on a 10% denaturing gel and RNAs isolated from gel slices. Partial alkaline hydrolysis of RNA was performed by incubating the RNA in 0.1 M $NaHCO_3$, pH 9.0, along with 20 μg of yeast tRNA, at 95 °C for 6 min; the RNAs were then precipitated with ethanol.

miRNA Expression from a Plasmid in Transfected Cells—DNA spanning the precursor miR-31 sequence was amplified from human genomic DNA (Clontech), digested with HindIII and XhoI, and then cloned into a modified pSuper vector (15). The constructs were then co-transfected with a plasmid expressing a control short hairpin RNA, specific for green fluorescent protein, into 293T cells; RNA was isolated 2 days later and analyzed by Northern blotting as described previously (15, 16).

RNA Binding to Glutathione S-Transferase (GST) Proteins—A DNA template for ssRNA substrate transcription was prepared by PCR amplification of a 43-bp multiple cloning site fragment derived from pCMV (17). The DNA template for precursor miR-30a transcription was prepared by annealing and extending the two oligonucleotides 5′-TGTA-ATACGACTCACTATAGGTAAACATCCTCGACTGGAAGCT-3′ and 5′-GCGGCAAACATCCGACTGAAAGCC<u>CATCTGTGGCTTCAC</u>AGCTTC CAGTCGAGG-3′ (the putative precursor miR-30a loop region is underlined). The DNA template for precursor miR-30a L5 transcription was prepared by annealing and extending the two oligonucleotides 5′-TGT-AATACGACTCACTATAGGTAAACATCCTCGACTGGAAGCT-3′ and 5′-GCGGCAAACATCCGACTGAAGCC<u>AGATT</u>AGCTTCCAGTCGAG-G-3′ (the putative loop region is underlined). The DNA template for precursor miR-30 L5+10 transcription was prepared by the same procedure using the two oligonucleotides 5′-TGTAATACGACTCACTATA-GGTAAACATCCTCGACTGGAAGCT-3′ and 5′-*ACCACCTAAC*GCGG-CAAACATCCGACTGAAAGCC<u>AGATT</u>AGCTTCCAGTCGAGG-3′ (the putative loop region is underlined, and the 10 nt 3′ extension is italicized). Other DNA templates and RNA probes were synthesized as described above.

For protein purification, DH5α cells transformed with pGEX-4T-1 or one of its derivatives was induced and lysed, and GST proteins were bound to glutathione beads (Amersham) as described previously (19). The beads were washed three times with 10 mM Tris, pH 7.6, 0.5 M LiCl, and 0.1% Triton X-100 followed by one wash with binding buffer (20 mM Tris, pH 7.6, 0.1 M KCl, 0.1% Tween 20, 0.1% Triton X-100); and the beads were then incubated with ~90 μl of binding buffer containing 40 units of RNasin, 10 mg/ml poly(dI-dC) (Sigma), and various ^{32}P-labeled RNA substrates (~10^4 cpm of each) at 4 °C for ~25 min. Beads were afterward washed four times with binding buffer. RNA was then eluted with 100 μl of 1% SDS, 0.15 M NaCl, and 30 μg of yeast tRNA, extracted with phenol/chloroform, and precipitated with ethanol. Bound RNA was analyzed by electrophoresis and autoradiography. In a parallel experiment, proteins bound to glutathione beads were analyzed directly by electrophoresis followed by Coomassie Blue staining to confirm that the relevant proteins were indeed purified. All of the experiments were performed at least two times with identical results.

RESULTS

Drosha-mediated Cleavage of pri-miRNAs in Vitro Requires Flanking ssRNA—We prepared FLAG epitope-tagged Drosha enzyme by transfecting human 293T cells with the plasmid pCK-Drosha-FLAG (11), purifying proteins with FLAG-agarose, and then eluting the immunoprecipitate with 3× FLAG peptide. The immunoprecipitate contained Drosha-FLAG and presumably also endogenous DGCR8 (4). For RNA substrates, we prepared transcripts encoding three human miRNAs, miR-31, miR-223, and miR-30a (miR-30 for short). ^{32}P-Labeled RNA substrates, including various flanking sequences beyond the precursor miRNA hairpin, were mixed with the Drosha immunoprecipitate and tested for the generation of the ~60-nt pre-miRNA intermediate (marked by *asterisks* in Figs. 1–6). Figs. 1–3 list all of the natural or near natural pri-miRNA variants tested along with their predicted secondary structures and also show some of the representative autoradiographs. RNA substrates were named according to the numbers of extra nucleotides 5′ and 3′ to the predicted pre-miRNA cleavage product, *e.g.* miR-31(13+16) denotes a miR-31 variant with 13 nt flanking the 5′ end and 16 nt flanking the 3′ end of the predicted precursor miR-31 hairpin (Fig. 1). The RNA located outside of a pre-miRNA hairpin can be tentatively demarcated into two distinct components: one is the essential stem extension located immediately adjacent to the pre-miRNA hairpin, and the other is the extra RNA at the ends that presumably forms ssRNA extensions. Importantly, we found that these predicted ssRNA flanking sequences greatly facilitated Drosha cleavage.

Fig. 1 presents the results using primary miR-31 variants. The miR-31(13+16) substrate was processed efficiently by Dro-

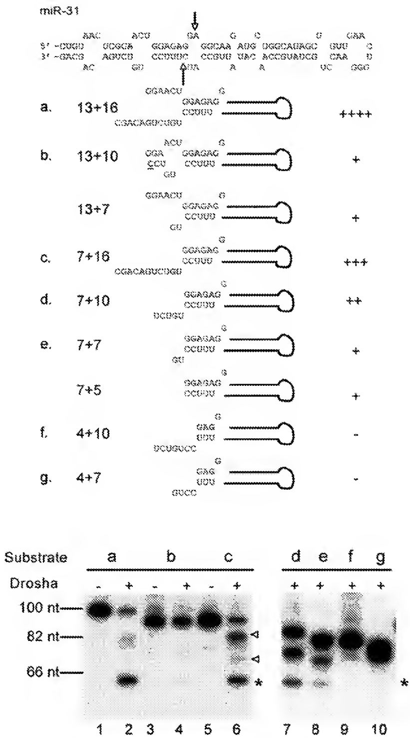


FIG. 1. **Schematic of miR-31 transcripts and their susceptibility to Drosha cleavage *in vitro*.** *Top*, *arrows* indicate the putative Drosha cleavage sites that liberate the precursor miR-31 RNA hairpin. The *hairpin symbol* below represents precursor miR-31. RNA variants are named based on the numbers of extra nucleotides 5′ and 3′ to the precursor miRNA. For example, *13+16* means that the variant has 13 nt extra on the 5′ side and 16 nt on the 3′ side, and *13+10* means that the variant has 13 nt extra on the 5′ side and 10 nt on the 3′ side. A residue different from the endogenous transcript is *underlined*. A series of + and − signs are used to denote the efficiencies with which Drosha cleaved its substrates: −, no cleavage at all; ++++, the highest level of processing. Reduced levels were judged based on reduced precursor miRNA production and increased intermediate accumulation. *Bottom*, RNA variants with their original *in vitro* processing data are indicated as substrates *a*, *b*, *c*, *d*, *e*, *f*, and *g*. The size markers to the *left* of the autoradiograph were DNAs. *Asterisks* indicate the position of precursor miR-31, and *arrowheads* point to processing intermediates.

sha, generating largely precursor miR-31 (Fig. 1, *lane 2*) and very few processing intermediates (indicated by *arrowheads*). Judged from their characteristic sizes, these intermediates were RNAs cut by Drosha at the authentic 5′ or 3′ cleavage site but not at both. Such products have been reported previously (4). These singly cut RNAs likely represent dead-end products *in vitro* because they lack an essential feature required for *de novo* processing (*i.e.* a stem extension beyond the Drosha cleavage sites; see below). When isolated from gels and treated with Drosha again, they were indeed totally resistant to cleavage (data not shown). Eliminating part of the 5′ flanking ssRNA, as in miR-31(7+16), led to the accumulation of singly cut intermediates (Fig. 1, *lane 6*). Further deletion of part of the 3′ ssRNA extension, as in miR-31(7+10) or miR-31(7+7), led to a further reduction in precursor miR-31 production (Fig. 1, *lanes 7* and *8*). When deletions were made even closer to the precursor miR-31 region, *e.g.* in miR-31(4+10) or (4+7), Drosha processing became undetectable (Fig. 1, *lanes 9* and *10*).

Very similar results were also obtained for miR-223 (Fig. 2) and miR-30 (Fig. 3). For example, compared with miR-223-(29+21), the 3′ ssRNA-shortened substrates miR-223(29+15) and miR-223(29+12) were much less efficiently cleaved by Drosha *in vitro*. Blunt-ended RNA hairpins, such as miR-31(7+5) (Fig. 1), miR-223(16+15) (Figs. 2 and 5, and see below), and miR-30(11+9) (Fig. 3, *lane 2*), were cleaved very poorly or not at all. In general, the longer the native flanking sequence the miRNA retained on both sides, the better substrate the RNA was for Drosha processing *in vitro*. However,

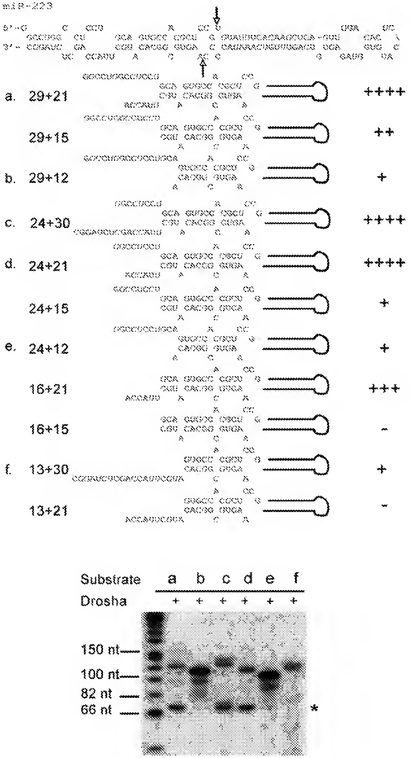


FIG. 2. **Drosha processing of miR-223 transcripts *in vitro*.** See Fig. 1 legend for the meaning of symbols and labels.

flanking ssRNA at the 3′ end did appear to be more critical than the one at the 5′ end. Thus, miR-223(16+21) was a better Drosha substrate than miR-223(24+15), although neither was as good as miR-223(24+21) (Fig. 2). Finally, and consistent with previous reports (11, 15, 16), a stem extension beyond the pre-miRNA hairpin, within the longer pri-miRNA, was always essential, although the exact length requirement varied among the three miRNAs tested. miR-223 apparently requires the longest extension, as miR-223(13+30) and miR-223(13+21) were hardly processed by Drosha (Fig. 2). In contrast, miR-31(7+16) and miR-31(7+10) were still reasonable Drosha substrates (Fig. 1, *lanes 6* and *7*), whereas further shortening of the stem extension to make miR-31(4+10) or miR-31(4+7) abolished cleavage (Fig. 1, *lanes 9* and *10*).

Flanking ssRNA Sequences Function in a Largely Sequence-independent Manner in Vitro—Figs. 1–3 show secondary structure predictions for miRNA hairpins based on MFOLD. The actual RNA folding details at the top of the hairpin and at the base of the stem and the conformations of the flanking RNAs might be dynamic and/or different from the RNA structures proposed here. For example, some residues from the 5′ ssRNA extension could potentially form hydrogen bonds with those from the 3′ side. Nevertheless, we hypothesized that Drosha preferred flanking RNA sequences that did not fold into a

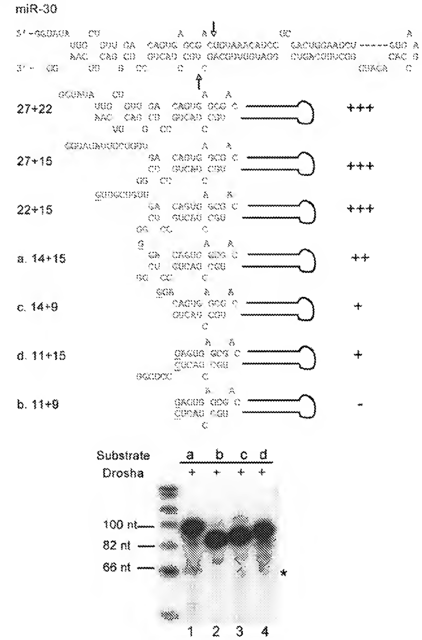


FIG. 3. **Drosha processing of miR-30 transcripts *in vitro*.** See Fig. 1 legend for explanation of symbols and labels. *Underlined letters* at the overhangs represent nucleotides different from the endogenous primary miR-30, and they were changed to fit the transcription start site used by T7 RNA polymerase or for maintaining secondary structures.

helical conformation. We performed several experiments to test this idea, and these results are shown in Figs. 4–6. In Fig. 4*A*, when the predicted 6-nt 5′ ssRNA extension and 11-nt 3′ extension in the natural but truncated miR-31(13+16) substrate were replaced by arbitrary sequences, the new RNA, called (6)+7+5+(11), was still effectively cleaved by Drosha *in vitro* (compare *lane 2* and *lane 6*). miR-223 is another example (Fig. 4*B*). Here, the artificial (8)+16+15+(6) variant was cleaved at an efficiency close to that of the natural miR-223(24+21) substrate.

To examine flanking sequence requirements in more detail, we turned to miR-223(16+21) (Fig. 5). This RNA is predicted to form a simple structure containing a 6-nt 3′ ssRNA overhang, and *in vitro* processing by Drosha was weaker than seen with miR-223(24+21) but still readily detectable (Fig. 5, compare *lane 2* with *lane 12*). Deleting the 6-nt 3′ overhang eliminated detectable Drosha processing (Fig. 5, *lane 4*), which was rescued by adding back 6 nt (*lane 6*) or 9 nt (*lane 8*) but not 3 nt (*lane 10*) of an arbitrary ssRNA sequence. This rescue was not simply due to larger RNA size because introducing a 6-nt 5′ extension that is predicted to form a 6-bp stem with the arbitrary 6-nt 3′ extension abolished processing (data not shown). Furthermore, when an artificial hairpin was appended to the 6-nt 3′ extension to make 16+15+(6+D), the new RNA was processed less efficiently than the parental 16+15+(6) RNA (Fig. 6, compare *lane 4* with *lane 2*), whereas an identically sized RNA substrate lacking the predicted hairpin structure, 16+15+(6+S), was processed more efficiently (Fig. 6, *lane 6*). From these results, we concluded that flanking ssRNA sequences strongly enhance Drosha cleavage of pri-miRNA hairpins but that the particular sequence of the ssRNA extensions was not critical as long as the flanking ssRNA sequences were of sufficient length (>3 nt) and adopted a largely single-stranded structure.

Free RNA Ends Are Not Required for Drosha Processing—If Drosha first recognizes an unpaired 5′ end and/or 3′ end and then scans along the RNA for a suitable stem-loop structure, such a mechanism could explain why a ssRNA extension is required. To test this scenario, we chose a primary miR-31 substrate that was larger than the ones examined above and

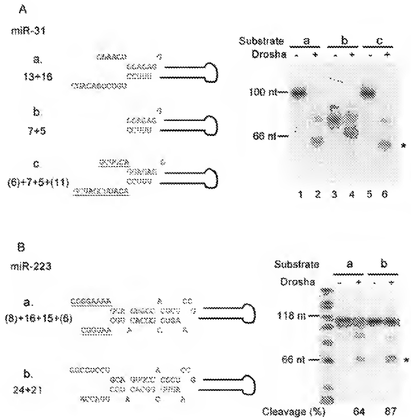


FIG. 4. **Sequence of the ssRNA extension can be replaced without significantly affecting cleavage.** *A*, the 6-nt 5′ overhang and 11-nt 3′ overhang of primary miR-31 (*13+16*) were substituted with arbitrary sequences (*underlined*) to make the (6)+7+5+(11) mutant. The *hairpin symbol* represents pre-miRNA. Three RNA substrates, *i.e.* 13+16, 7+5, and (6)+7+5+(11) (substrates *a*, *b*, and *c*) were subjected to *in vitro* processing by Drosha, as shown on the *right*. Size of DNA markers is shown to the *left* of the autoradiograph, and the *asterisk* indicates the pre-miRNA band. *B*, the native primary miR-223 (*24+21*) overhangs were likewise replaced to make (8)+16+15+(6). Labels are the same as in *A*. Cleavage percentage was calculated as the ratio of the intensity of the pre-miRNA band (marked by an *asterisk*) divided by that of the remaining full-length substrate and corrected for cytosine contents.

contained an ~85-nt sequence flanking each side of the mature precursor miR-31. The linear RNA was a good substrate for Drosha processing *in vitro* (Fig. 7*B*, *lane 2*) and was properly processed when transcribed in transfected cells (15). We prepared its closed, circular version as described under 'Materials and Methods' (Fig. 7*A*, *lane 2*). The identity of the circular RNA was confirmed by partial alkaline hydrolysis as this treatment collapsed it to a position corresponding to the linear RNA and yielded a smear below it (Fig. 7*B*, compare *lanes 3* and *6*). Drosha cleaved this circular RNA efficiently to generate precursor miR-31 and the predicted single by-product (Fig. 7*B*, *lane 5*). Thus, free RNA ends are not necessary for Drosha processing of pri-miRNA hairpins *in vitro*.

miRNA Expression in Cells Requires ssRNA Sequences Flanking the pri-miRNA Hairpin—We next asked whether RNA variants that were good substrates for Drosha cleavage *in vitro* were also good substrates for processing to a mature miRNA in cells. We inserted the corresponding DNAs behind an RNA polymerase III-dependent promoter, the H1 promoter, transfected the resultant expression plasmids into human 293T cells, and then examined miRNA expression by Northern blotting. The vector that we used (15) contributed ~8 nt at each side of the cloned RNAs (Fig. 8). We found that even longer sequences were required for, or at least enhanced, miRNA maturation in this setting. As shown in Fig. 8, miR-31(51+51), which had 51 nt of natural RNA sequence flanking each side of precursor miR-31 (substrate *a*), yielded a high level of mature miR-31; miR-31(51+16) and miR-31(13+51) had reduced levels of miR-31 production (substrates *b* and c); and miR-31(13+16) gave only small amounts of mature miR-31 (substrate *d*). We then added artificial sequences to this shorter pri-miRNA transcript to bring it back to the size of miR-31(51+51). These sequences were designed so that they were clearly different from the natural ones and also so that they would not form a strong secondary structure. Adding an arbitrary 38-nt sequence at the 5′ side of miR-31(13+51) to make miR-31-(38+13+51) largely restored miR-31 expression (Fig. 8, compare substrates c and *f*). However, making miR-31(51+16+35) and miR-31(38+13+16+35) only partially restored miR-31 expression (Fig. 8, substrates *e* and *g*). The failure to completely rescue miR-31 expression by adding back artificial sequences in this latter instance could be due to the loss of a positive contribution from the natural flanking sequences or to the introduction of negative effects by the new flanking ssRNAs.

The R-rich Region of the Drosha Subunit Preferentially Binds ssRNA—We have shown previously that Drosha prefers to process pri-miRNA hairpins bearing a large ssRNA terminal loop (15), and here, we have further demonstrated that ssRNA extensions are required for Drosha cleavage *in vitro*. To identify which part(s) of the Drosha-DGCR8 complex interacts with ssRNA, we expressed and purified individual domains of the Drosha and DGCR8 subunits as GST fusion proteins in *E. coli* and tested their interaction with various RNA substrates. Fig. 9 presents the results obtained using the dsRBDs of Drosha and DGCR8 and the R-rich domain of Drosha. Fig. 9*A* shows the domain structures of the proteins, and Fig. 9*B* lists the different RNA substrates used in the binding experiments. Substrate *a* (Fig. 9*B*) is a 43-nt RNA derived from a vector sequence and is used here as a representative of ssRNA. MFOLD predicts that it contains no consecutive helical RNA region longer than 5 bp. Substrate *b* (Fig. 9*B*) is the native pre-miRNA for miR-30, which is a 63-nt hairpin bearing a 2-nt 3′ overhang (11). Substrate *d* (Fig. 9*B*), the L5 variant, is similar to substrate *b*, but because of substitution (15), it contains a small 5-nt terminal loop instead of the predicted 15-nt loop (see 'Materials and Methods'). Compared with wild-type primary miR-30, a pri-miRNA bearing the L5 mutant is processed much less efficiently by Drosha (15), underscoring the importance of a large terminal loop. Substrate *e* (Fig. 9*B*), L5+10, has 10 nt of arbitrary ssRNA sequence appended to the original 3′ overhang of substrate *d*. Substrates *c* and *f* (Fig. 9*B*) are the same primary miR-223 RNA substrates listed in Fig. 2.

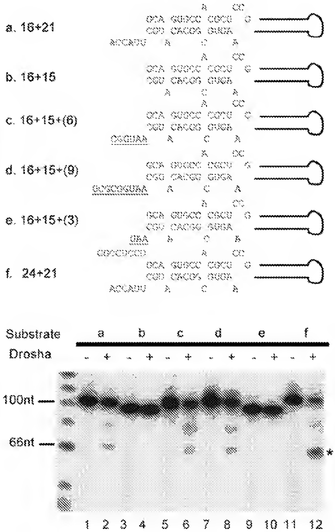


FIG. 5. **Size requirement for the 3′ overhang of the primary miR-223 transcript.** Six RNA substrates (*a–f*) containing a 6-nt natural 3′ overhang, no overhang, 6, 9, or 3 nt of an arbitrary 3′ ssRNA overhang (*underlined*), or extensions at both the 5′ and 3′ sides were examined for Drosha processing *in vitro*. Labeling is the same as in Fig. 4.

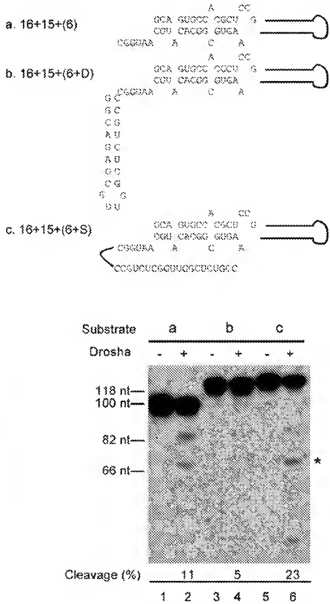


FIG. 6. **Drosha cleavage of miR-223 substrates with different 3′ extensions.** As shown, substrate *b*, *16+15+(6+D)*, has a predicted 8-bp stem and a 4-nt loop flanking the pri-miRNA hairpin, whereas substrate *c*, *16+15+(6+S)*, is predicted to have a single-stranded 3′ extension. Cleavage efficiencies were calculated and labeled the same as in Fig. 4*B*.

All of the Drosha and DGCR8 protein fragments were expressed as GST fusions in bacteria (Fig. 9*D*), with the Drosha RBD and the DGCR8 2XRBDs being expressed at a higher level than the level in the R-rich region of Drosha. GST-DGCR8 2XRBDs bound avidly to all of the RNAs tested (Fig. 9*C*, *lanes 7* and *14*). Single, individual dsRBDs of DGCR8 bound RNA as well as did the 2XRBDs (data not shown). The GST-Drosha R-rich domain fusion, however, pulled down little or no substrate *d* (Fig. 9*C*, *lane 13*, L5 RNA). The GST-Drosha R-rich domain fusion did, however, interact very well with all other substrates tested, such as the putative ssRNA (Fig. 9*C*, substrate *a*) and hairpin RNAs with a wild-type, presumably large and flexible, terminal loop (substrates *b*, *c*, and *f*) or a 3′ ssRNA extension, as in substrates *e* and *f*. GST alone or GST-Drosha RBD did not bind any RNA under these conditions, and other Drosha truncations were expressed too poorly in *E. coli* for us to test their RNA binding potential. Curiously, if poly(dI-dC) was omitted in the binding reactions, Drosha RBD could then exhibit RNA binding activity (data not shown).

DISCUSSION

The principal finding of our current study is that efficient Drosha processing of a pri-miRNA substrate *in vitro* needs a substantial ssRNA flanking sequence attached to an extensive, ~80-nt pri-miRNA stem-loop structure. Such a requirement for flanking ssRNA sequences has not been observed for other RNase III type enzymes. *E. coli* RNase III, yeast Rnt1p, and PacI can all cleave dsRNAs with 5′ and 3′ ends arranged in base pairs (10, 20). Although Dicer clearly prefers a 2-nt 3′ overhang, human Dicer nevertheless cleaves dsRNAs containing a blunt end with only a small drop in efficiency *in vitro* (3, 7). Even allowing for some uncertainty at the terminal structure, such as breathing, our data indicate that Drosha requires much longer ssRNA sequences flanking its dsRNA substrate than do other RNase III enzymes.

Drosha does not cleave a fully helical RNA (21), thus suggesting that ssRNA is involved in mediating this protein-RNA interaction. We showed previously that Drosha strongly prefers a large, unstructured terminal loop on its pri-miRNA substrates (15), which together with the essential ssRNA overhangs identified in this study may thus satisfy the predicted

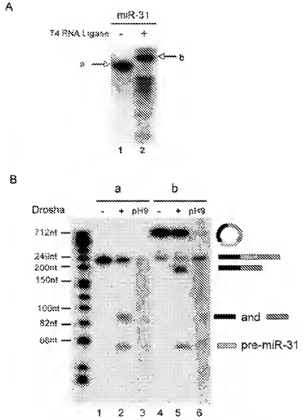


FIG. 7. **Processing of linear and circular substrates.** *A*, a linear primary miR-31 transcript and a self-ligated derivative were fractionated on a 10% denaturing gel. *Bands a* and *b*, indicated by *arrows*, were excised from the gel and the RNA eluted. *B*, RNAs from *bands a* and *b* in *A* were subjected to Drosha cleavage *in vitro* (*lanes 2* and *5*) or to partial alkaline hydrolysis (*lanes 3* and *6*). Different predicted RNA species are *diagrammed* at the *right*. Precursor miR-31 (*pre-miR-31*) is represented by a *light gray bar*, and the flanking sequences by *dark gray* and *black bars*.

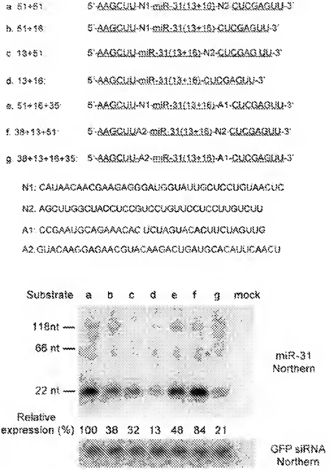


FIG. 8. **Mature miR-31 expression in transfected cells.** The primary miR-31 substrates (*a–g*) were transcribed from plasmids and contain different sequences flanking the predicted precursor miR-31 intermediate. *Underlined* are the elements common to all of the substrates. Sequences at the 5′ and 3′ ends are from the vector. *N1* and *N2* represent natural sequences flanking miR-31(13+16) (Fig. 1). *A1* and *A2* represent arbitrary sequences. The *bottom panels* show the results of a Northern analyses for miR-31, with a green fluorescent protein small interfering RNA (*GFP siRNA*) used as control, using RNA derived from transfected 293T cells. Sizes of DNA markers are shown at the *left*. Relative expression was calculated as the mature miR-31 signal divided by that of the green fluorescent protein small interfering RNA signal, with cleavage of substrate *a* set as 100%.

requirement for single-stranded RNA for Drosha cleavage. We believe that we have now largely defined the RNA elements that are essential for Drosha cleavage *in vitro*, and it is apparent that Drosha actually engages a very large RNA surface. Unlike other RNase III enzymes that can function alone and recognize smaller and simpler RNA structures, at least *in vitro*, pri-miRNA processing is actually mediated by a protein complex that minimally consists of the catalytic subunit Drosha and a protein partner called DGCR8 in humans or Pasha in invertebrates (4, 12, 13). DGCR8 greatly enhances, and is likely indeed necessary for, Drosha activity. There is little information as to how these proteins interact with their RNA substrates. As the first step toward achieving such an understanding, we show here that, under our assay conditions, the dsRBDs of DGCR8 were capable of binding both to RNA with a largely single-stranded conformation and to RNA with a mostly helical structure (Fig. 9). Although the dsRBD of Drosha showed a very low affinity for RNA, interestingly we found that the R-rich region of the Drosha subunit had a preference for ssRNA. For the R-rich region, the ssRNA can be either at the top of the stem, *i.e.* in the terminal loop, or flanking the base of the RNA hairpin (Fig. 9). A 2-nt 3′ overhang together with a small terminal loop, as seen in the precursor miR-30 L5 variant, is insufficient to support binding. The Drosha R-rich region is not the sole determinant on the enzyme that requires ssRNA, as it does not distinguish between miR-223(24+21), a good substrate for Drosha cleavage *in vitro*, and miR-223-(16+15), in which cleavage by Drosha was never observed. Contributions to recognition of ssRNA regions within the pri-miRNA substrate may thus also come from other, as yet undefined, parts of the Drosha-DGCR8 complex.

A recent article (4) indicated that a Drosha mutant devoid of the R-rich region analyzed here could still function as an active pri-miRNA processing enzyme when transiently expressed in cells. The Drosha deletion mutant was expressed at a much higher level than the full-length protein (4), so it will be interesting to see whether it indeed retains the same specific cleavage activity and substrate specificity as the full-length enzyme. Because Drosha can self-associate, it is formally possible that endogenous Drosha protein could form a complex with the over-expressed Drosha deletion mutant and then exhibit cleavage activity *in vitro*. Our data do not address the question of what role the R-rich domain plays in an intact Drosha protein in cells, and the mechanisms governing Drosha-RNA interactions certainly need to be investigated further.

For many pri-miRNAs, RNA folding algorithms predict that sequences at the 5′ side and the 3′ side, beyond the pre-miRNA hairpin, can anneal to form a very long, imperfect stem. A modest stem extension adjacent to the pre-miRNA intermedi-

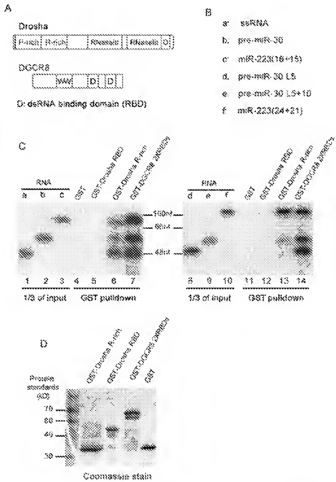


FIG. 9. **RNA binding to individual domains of Drosha and DGCR8.** *A*, domain structure of the Drosha and DGCR8 subunits. The R-rich region and the dsRBD domain of Drosha were separately expressed as GST fusions in bacteria. Part of the DGCR8 from glutamic acid 502 to the end of the protein, which includes the two dsRBDs (*D*), was also expressed as a GST fusion in bacteria. *B*, list of the RNA substrates (*a–f*) that were tested for protein binding in *C*. For predicted RNA structure details, see 'Materials and Methods' and Fig. 2. *C*, RNA binding to GST proteins immobilized on glutathione beads. *Lanes 1*, *2*, *3*, *8*, *9*, and *10* show approximately one-third of the total input RNA. *Lanes 4–7* and *11–14* show RNAs bound by recombinant proteins *in vitro*. For *lanes 4–7*, RNA substrates *a*, *b*, and c were mixed in a single solution and incubated with the proteins. For *lanes 11–14*, RNA substrates *d*, *e*, and *f* were also mixed and incubated with the proteins. Similar results were obtained when only a single RNA was used for binding. *D*, Coomassie staining of a protein gel to show that the expected GST fusion proteins were indeed made. *P-rich*, proline-rich; *pre-miR-30*, precursor miR-30.

ate is indeed essential for the excision of the pre-miRNA intermediate from a pri-miRNA substrate, but a longer dsRNA conformation is not beneficial and can be inhibitory (Refs. 11, 15, and 16 and this study). The exact sequence of the ssRNA extension is apparently not critical, but a strong RNA secondary structure within the flanking sequence or formed between the 5′ and 3′ extensions is distinctively disfavored (Figs. 4–6). Although it is currently unclear as to why ssRNA extensions are needed or how they regulate Drosha function, we favor the hypothesis that the flanking ssRNA sequences form part of the Drosha-RNA interface, *e.g.* Drosha may simultaneously bind to the stem-loop structure as well as to the overhang(s). Alternatively, Drosha may be intrinsically unable to bind or correctly position itself directly onto a hairpin structure. It is conceivable that the extra flanking sequences may be needed initially to tether or recruit the Drosha-DGCR8 complex to RNA. In the absence of a suitable stem-loop structure nearby, however, the enzyme may rapidly dissociate from ssRNA binding sites. Because Drosha is capable of cleaving a circular substrate (Fig. 7), we can exclude the possibility that the RNA overhang contributes a free 5′ or 3′ end necessary for Drosha function. Considering that many miRNAs are encoded within the introns of their host genes (22), our data are consistent with the prediction that Drosha can operate directly on lariat RNAs. It is also, of course, possible that Drosha cleavage *in vivo* might be facilitated by other, as yet unknown, proteins.

We found that the minimal RNA element required for *in vitro* Drosha cleavage, identified here as an ~80-nt RNA hairpin structure plus ~10-nt ssRNA overhang(s), was ineffective in mature miRNA production when transcribed from the H1 promoter in transfected cells (Fig. 8). This is consistent with an earlier report that at least 40 nt of extra sequences on each side of the pre-miRNA hairpin are required for efficient miRNA production in cells (18). Part of these 40 nt of extra RNA would form the stem and the ssRNA extensions essential for Drosha recognition and cleavage *in vitro* (11, 15, 16), but how the additional nucleotides contribute to pri-miRNA processing remains unknown. Drosha might need an even larger RNA structure for cleavage *in vivo*. Alternatively, the extra RNA sequences might affect transcription, RNA folding and/or RNA stability. Most miRNAs are transcribed from RNA polymerase II promoters *in vivo*, so it is also possible that transcription from the H1 promoter, an RNA polymerase III promoter, can inhibit Drosha function. A more detailed analysis of sequence requirements for miRNA processing *in vivo* will be required to fully address these questions.

Acknowledgments—We thank N. Kim and R. Shiekhattar for providing FLAG-tagged Drosha and DGCR8 expression plasmids, respectively.

REFERENCES

1. Lamontagne, B., Larose, S., Boulanger, J., and Abou Elela, S. (2001) *Curr. Issues Mol. Biol.* **3,** 71–78
2. Nicholson, A. W. (2003) in *RNAi: A Guide to Gene Silencing* (Hannon, G. J., ed) pp. 149–174, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
3. Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E., and Filipowicz, W. (2004) *Cell* **118,** 57–68
4. Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., and Kim, V. N. (2004) *Genes Dev.* **18,** 3016–3027
5. Zhang, K., and Nicholson, A. W. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94,** 13437–13441
6. Provost, P., Dishart, D., Doucet, J., Frendewey, D., Samuelsson, B., and Radmark, O. (2002) *EMBO J.* **21,** 5864–5874
7. Zhang, H., Kolb, F. A., Brondani, V., Billy, E., and Filipowicz, W. (2002) *EMBO J.* **21,** 5875–5885
8. Chanfreau, G., Buckle, M., and Jacquier, A. (2000) *Proc. Natl. Acad. Sci. U. S. A.* **97,** 3142–3147
9. Nagel, R., and Ares, M., Jr. (2000) *RNA (N. Y.)* **6,** 1142–1156
10. Lamontagne, B., and Abou Elela, S. (2004) *J. Biol. Chem.* **279,** 2231–2241
11. Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. (2003) *Nature* **425,** 415–419
12. Denli, A. M., Tops, B., Plasterk, R. H. A., Ketting, R. F., and Hannon, G. J. (2004) *Nature* **432,** 231–235
13. Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004) *Nature*, **432,** 235–240
14. Bartel, D. P. (2004) *Cell*, **116,** 281–297
15. Zeng, Y., Yi, R., and Cullen, B. R. (2005) *EMBO J.* **24,** 138–148
16. Zeng, Y., and Cullen, B. R. (2003) *RNA (N. Y.)* **9,** 112–123
17. Zeng, Y., Wagner, E. J., and Cullen, B. R. (2002) *Mol. Cell* **9,** 1327–1333
18. Chen, C. Z., Li, L., Lodish, H. F., and Bartel, D. P. (2004) *Science* **303,** 83–86
19. Frangioni, J. V., and Neel, B. G. (1993) *Anal. Biochem.* **210,** 179–187
20. Lamontagne, B., Ghazal, G., Lebars, I., Yoshizawa, S., Fourmy, D, and Abou Elela, S. (2003) *J. Mol. Biol.* **327,** 985–1000
21. Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001) *Nature* **409,** 363–366
22. Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., and Bradley, A. (2004) *Genome Res.* **14,** 1902–1910

The EMBO Journal Vol.16 No.22 pp.6636–6645, 1997

Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin

Bruno Martoglio[1], Roland Graf[2] and Bernhard Dobberstein

Zentrum für Molekulare Biologie der Universität Heidelberg (ZMBH), Postfach 106249, 69052 Heidelberg, Germany and [2]Laboratorium für Biochemie II, ETH-Zentrum, Universitätstrasse 16, 8092 Zürich, Switzerland

[1]Corresponding author

Secretory proteins and most membrane proteins are synthesized with a signal sequence that is usually cleaved from the nascent polypeptide during transport into the lumen of the endoplasmic reticulum. Using site-specific photo-crosslinking we have followed the fate of the signal sequence of preprolactin in a cell-free system. This signal sequence has an unusually long hydrophilic n-region containing several positively charged amino acid residues. We found that after cleavage by signal peptidase the signal sequence is in contact with lipids and subunits of the signal peptidase complex. The cleaved signal sequence is processed further and an N-terminal fragment is released into the cytosol. This signal peptide fragment was found to interact efficiently with calmodulin. Similar to preprolactin, the signal sequence of the HIV-1 envelope protein p-gp160 has the characteristic feature for calmodulin binding in its n-region. We found that a signal peptide fragment of p-gp160 was released into the cytosol and interacts with calmodulin. Our results suggest that signal peptide fragments of some cellular and viral proteins can interact with cytosolic target molecules. The functional consequences of such interactions remain to be established. However, our data suggest that signal sequences may be functionally more versatile than anticipated up to now.

Keywords: calmodulin/eukaryotic signal peptidase/HIV-1 gp160/prolactin secretion/signal sequence

Introduction

Secretory proteins and membrane proteins contain a signal sequence for targeting to the protein-conducting channel and subsequent translocation across or insertion into the endoplasmic reticulum (ER) membrane (Rapoport *et al.*, 1996). During transport into the ER lumen the signal sequence is often cleaved from the precursor protein by the signal peptidase (Blobel and Dobberstein, 1975). The characteristic feature of a signal sequence is a tripartite structure: a polar N-terminal n-region, a hydrophobic core (h-region) of 7–15 residues and a polar C-terminal c-region that contains the consensus sequence for signal peptide cleavage (von Heijne, 1985). The n-region of most signal sequences comprises only a few residues. However, some signal sequences have extended n-regions, of up to 150 residues. The function of such long n-regions is not as yet known.

The fate of only a few signal sequences has been elucidated. Fragments derived from the signal sequence of some secretory proteins or type I membrane proteins have been found associated with MHC class I molecules and are transported to the cell surface for presentation to cytolytic T cells. Some signal peptide fragments (SPFs) corresponding mainly to C-terminal segments of the respective signal sequence become associated with MHC class I molecules independent of the transporters associated with antigen processing (TAP) (Henderson *et al.*, 1992; Wei and Cresswell, 1992). However, for one SPF derived from the n-region of the lymphocytic choriomeningitis virus envelope protein a strictly TAP-dependent binding to MHC class I molecules has been reported (Hombach *et al.*, 1995). These results indicate that SPFs can be released from the membrane to the ER lumen or to the cytosol. Besides functioning in antigen presentation, nothing is known about the physiological roles of SPFs released into the cytosol or the ER lumen.

Using a synchronized *in vitro* system we have previously shown that the cleaved signal peptide of the secretory protein hormone preprolactin (p-Prl) is further processed in the ER membrane and that the resulting N-terminal SPF is released into the cytosol (Lyko *et al.*, 1995). Processing of the cleaved signal sequence was found to be sensitive to the immunosuppressive proline isomerase inhibitor cyclosporin A (Klappa *et al.*, 1996). Cyclosporin A is known to bind to cellular proteins termed cyclophilins which have proline isomerase activity and are thought to modulate the activity of various enzymes (Schreiber and Crabtree, 1992). It is thus conceivable that a cyclophilin in the ER regulates signal sequence processing and subsequent release of the SPF into the cytosol.

To determine possible functions of SPFs released into the cytosol, we followed the fate of the p-Prl signal sequence and identified components interacting with the cleaved signal sequences in the membrane and the SPF in the cytosol using site-specific photo-crosslinking (Martoglio and Dobberstein, 1996). We found that in the cytosol the p-Prl SPF interacts efficiently with calmodulin (CaM). The p-Prl signal sequence has an extended basic n-region such that it can potentially form a basic amphipathic α(baa)-helix. This feature is characteristic for CaM binding domains (O'Neil and DeGrado, 1990) but is not found in the majority of signal sequences. The HIV-1 envelope protein gp160 also has a signal sequence with an extended n-region that can potentially form a baa helix. As with p-Prl, we followed the fate of the p-gp160 signal sequence and found that a p-gp160 SPF is released into the cytosol and interacts with CaM. A synthetic p-gp160 SPF corresponding to the N-terminal 23 amino acid residues of the p-gp160 signal sequence has high affinity

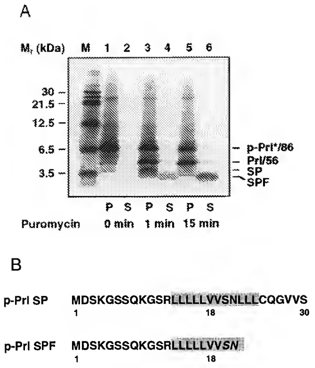


Fig. 1. Signal sequence cleavage, processing and release. (**A**) p-Prl*/86 chains were inserted into rough microsomes (lanes 1 and 2) and subsequently released from the ribosome by addition of puromycin. Incubation was continued for 1 (lanes 3 and 4) and 15 min (lanes 5 and 6). Before application to SDS–polyacrylamide gels membranes (pellet, P) were separated from the cytosol (supernatant, S) by centrifugation. SP indicates the cleaved signal sequence, SPF the signal peptide fragment. (**B**) Outline of the p-Prl signal sequence and the p-Prl SPF. The h-region of the signal sequence is shaded. Italic letters in the p-Prl SPF indicate that the C-terminal end of the fragment is estimated (see Lyko *et al.*, 1995).

for CaM and efficiently inhibits Ca^{2+}/CaM-dependent phosphodiesterase *in vitro*. Our results suggest that SPFs of distinct signal sequences may interfere with CaM functions and may act as regulatory peptides.

Results

Cleavage and processing of the p-Prl signal sequence

We used a previously established system to follow the fate of the cleaved signal sequence of p-Prl (Lyko *et al.*, 1995). A truncated mRNA coding for the 86 N-terminal amino acid residues of p-Prl was translated in the presence of rough microsomal membranes. Since truncated mRNAs lack a stop codon, termination of translation does not occur. Under these conditions p-Prl/86 chains are inserted into the translocation complexes of the microsomes and remain bound to the ribosome (Gilmore *et al.*, 1991). Signal sequence cleavage does not occur because the p-Prl/86 chains are too short (Figure 1A, lane 1). To remove non-inserted p-Prl/86 chains, microsomes are then isolated, resuspended in cytosolic extracts and p-Prl/86 chains released from the ribosome by addition of puromycin. p-Prl/86 chains become translocated across the microsomal membrane and the signal sequence is cleaved. Finally, membranes are separated from the cytosol by centrifugation and both the membrane pellet and the cytosol (supernatant) are analysed by SDS–PAGE.

When p-Prl/86 chains were released from the ribosome with puromycin and incubation was continued for 1 min, the signal sequence was cleaved by the signal peptidase and was found associated with the membrane (stage I; Figure 1A, lanes 3 and 4, and B). After longer incubation (15 min) the signal sequence was cleaved and processed further by an as yet unknown signal peptide peptidase and an SPF was found in the cytosol fraction (stage II; Figure 1A, lanes 5 and 6, and B). The same result was obtained when mRNA coding for a mutant p-Prl, p-Prl*, was used which contains additional methionines at positions 12 and 13 for better labelling of the SPF with [^{35}S]methionine (Lyko *et al.*, 1995).

Membrane components interacting with the cleaved p-Prl signal sequence

In order to probe the molecular environment of the signal sequence by site-specific photo-crosslinking, the photo-activatable amino acid L-4′-(3-[trifluoromethyl]-3*H*-diazirin-3-yl)phenylalanine [(Tmd)Phe; Figure 2A] was co-translationally incorporated into the p-Prl signal sequence instead of Val18 (see Figure 1B) to give p-Prl*^{T} (for site-specific photo-crosslinking using (Tmd)Phe see Martoglio and Dobberstein, 1996, and references therein). p-Prl*^{T}/86 chains were then used for membrane insertion and puromycin release as described above.

We first probed the molecular environment of the signal sequence at stage I, when the cleaved signal sequence is still associated with the membrane (see above). Crosslinking was induced with UV light 1 min after addition of puromycin. We found two major crosslink products with apparent molecular weights of ~20 kDa and 4–5 kDa in the membrane fraction as revealed by SDS–PAGE (Figure 2B, lane 3, arrow and star). Immunoprecipitations with antibodies directed against the n-region of the p-Prl signal sequence and against prolactin (Prl) respectively showed that both crosslink products contain the cleaved signal sequence but not the Prl portion (Prl56; Figure 2B, lanes 5 and 6). Thus, the cleaved signal sequence (30 residues, ~3–4 kDa) is crosslinked to components with estimated molecular weights of ~17 and ~1 kDa.

Signal sequences are cleaved from the nascent precursor protein by the signal peptidase. We therefore assumed that the cleaved signal sequence is in contact with subunits of the pentameric signal peptidase complex (SPC) having molecular weights of 12, 18, 21, 22/23 and 25 kDa (Evans *et al.*, 1986). Using antibodies against the four smaller SPC subunits we could immunoprecipitate the ~20 kDa crosslink product with anti-SPC21 and to a minor extent also with anti-SPC18 antibodies (Figure 2B, lanes 7 and 8, arrow), but not with anti-SPC12 and anti-SPC22/23 antibodies (not shown). Sequence analysis of SPC21 and SPC18 has shown that these two subunits are putative serine proteases with homology to SEC11, an essential component of the signal peptidase complex in yeast (Böhni *et al.*, 1988; Greenburg *et al.*, 1989; Shelness and Blobel, 1990).

We have previously reported that the signal sequence of p-Prl is in contact with lipid molecules when short p-Prl chains are inserted into the protein-conducting channel and the signal sequence is still attached to the precursor protein (Martoglio *et al.*, 1995). Based on this finding and judged by the size of the small molecule (~1 kDa) crosslinked to

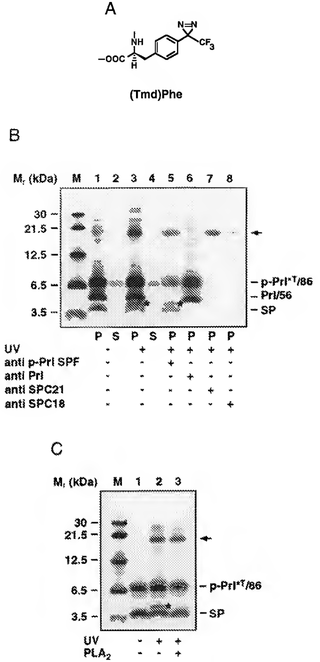


Fig. 2. Characterization of membrane components interacting with the cleaved p-Prl*^{T} signal sequence. (**A**) Schematic illustration of the photo-activatable amino acid L-4′-(3-[trifluoromethyl]-3*H*-diazirin-3-yl)phenylalanine [(Tmd)Phe], which was site-specifically incorporated at position 18 of the p-Prl signal sequence (see also Figure 1B). (**B**) Photo-crosslinking of the cleaved p-Prl*^{T} signal sequence to membrane components. p-Prl*^{T}/86 chains were released from the ribosome by puromycin and incubated for 1 min at 22°C. Samples were then frozen in liquid nitrogen and subjected to UV light (lanes 2–6). Membranes (P) were then separated from the cytosol (S) by centrifugation and analysed for crosslink products (lanes 1–4) or immunoprecipitated with antibodies directed against the p-Prl SPF (lane 5), Prl (lane 6) and subunits of the signal peptidase complex (SPC21, lane 7, and SPC18, lane 8) respectively. The arrow indicates crosslinks to SPC21 and SPC18. Stars indicate the small crosslink product. (**C**) Identification of phospholipid as crosslink partner. Membranes were treated with phospholipase A_2 (lane 3) after UV irradiation and separation from the cytosol. Samples were immunoprecipitated with anti-p-Prl SPF antibodies. The arrow indicates crosslinks to SPC 21 and SPC18, the star crosslinks to phospholipids.

the cleaved signal sequence, we expected the 4–5 kDa crosslink product shown in Figure 2B (lanes 3 and 5, star) to be a lipid adduct. To test whether the low molecular weight crosslink partner is a phospholipid, we treated the sample after crosslinking with bee venom phospholipase A_2 (Martoglio *et al.*, 1995). Phospholipase A_2 cleaves phospholipids at position C-2 into fatty acid and lysophospholipid. Because the amount of the 4–5 kDa crosslink product (Figure 2C, lane 2, star) was significantly reduced after treatment with phospholipase (Figure 2C, lane 3), we can conclude that a phospholipid is part of the respective crosslink product and hence that the cleaved signal sequence is also in contact with lipid molecules in the ER membrane.

Interaction of the p-Prl SPF with a cytosolic protein

We next probed the molecular environment of the signal sequence at stage II (see above). At this stage the p-Prl signal sequence has been cleaved and processed and an N-terminal SPF has been released into the cytosol (Figure 1A, lanes 5 and 6; Lyko *et al.*, 1995). Crosslinking was now induced 15 min after addition of puromycin. In the sample subjected to UV light we found the SPF and a crosslink product with an apparent molecular weight of ~20 kDa in the cytosol fraction (Figure 3A, lane 4). Immunoprecipitations with antibodies directed against the n-region of the p-Prl signal sequence (Figure 3A, lanes 5 and 6) and against Prl (Figure 3A, lanes 7 and 8) showed that the cytosolic crosslink product contains the SPF but not the Prl portion. Thus, the SPF is crosslinked to a component with an estimated molecular weight of 16–18 kDa (~20 kDa minus ~3 kDa from the SPF).

The cytosol we used for the experiments shown in Figure 3 was prepared either from bovine brain (Figure 3A, lanes 1–8) or GH_3 cells, a prolactin-synthesizing rat pituitary cell line (Figure 3A, lanes 11 and 12). When cytosol was omitted, no ~20 kDa crosslink product was obtained in the 'cytosol' fraction (Figure 3A, lanes 9 and 10). We have also tested cytosol prepared from a human cell line (Mel Juso cells; Figure 3B, lanes 5 and 6) as well as wheatgerm extract (Figure 3B, lanes 7 and 8). As shown in Figure 3B, the ~20 kDa crosslink product is always found when cytosol is present. This result suggests that the p-Prl SPF interacts with a cytosolic component uniformly present in higher eukaryotes.

The p-Prl SPF interacts with calmodulin

When p-Prl*^{T}/86 chains were released from the ribosome with EDTA instead of puromycin, the cytosolic ~20 kDa crosslink product was not found (Figure 4, lanes 1 and 2). This result suggests that release of the SPF and its binding to the cytosolic component depends on divalent cations. To test whether Ca^{2+} or Mg^{2+} is essential for binding, p-Prl*^{T}/86 chains were released from the ribosome with puromycin in the presence of EGTA to chelate calcium ions. Again, no cytosolic crosslink product was observed (Figure 4, lanes 3 and 4). This suggests a calcium dependence of SPF binding to a cytosolic component.

The estimated molecular weight of the cytosolic component that is crosslinked with the p-Prl SPF is 16–18 kDa. Calmodulin (CaM) is a cytosolic calcium binding protein of ~17 kDa and a central regulator of

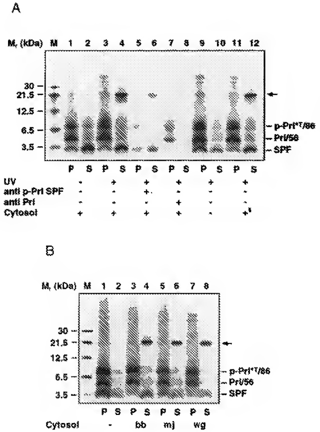


Fig. 3. The p-Prl*^{T} SPF is crosslinked to a component present in cytosol. (**A**) Photo-crosslinking of the p-Prl*^{T} SPF to a cytosolic protein. p-Prl*^{T}/86 chains were released from the ribosome by puromycin in the presence of cytosol prepared from bovine brain (lanes 1–8) or GH_3 pituitary cells (§, lanes 11 and 12). Cytosol was omitted in lanes 9 and 10. Samples were incubated for 15 min at 22°C and subjected to UV light (lanes 3–12) and membranes (P) separated from soluble components (S) by centrifugation. Samples were then analysed for crosslink products or immunoprecipitated with antibodies directed against the p-Prl SPF (lanes 5 and 6) and Prl (lanes 7 and 8) respectively. The major crosslink product is indicated by an arrow. (**B**) Photo-crosslinking of the p-Prl*^{T} SPF in cytosol prepared from various sources. p-Prl*^{T}/86 chains were released by puromycin as in (A) in the presence of cytosol prepared from bovine brain (bb, lanes 3 and 4), Mel Juso cells (mj, lanes 5 and 6) or wheatgerm extract (wg, lanes 7 and 8). Samples were further treated as in (A). The arrow indicates the major crosslink product in the cytosol fraction.

many kinases, phosphatases and transporters (Klee and Vanaman, 1982). To test whether the released p-Prl*^{T} SPF interacts with CaM, the potent CaM antagonist calmidazol was added to the crosslinking assay. In the presence of calmidazol the cytosolic ~20 kDa crosslink product was not observed (Figure 4, lanes 5 and 6), suggesting that calmidazol efficiently competes with the p-Prl SPF for CaM.

As shown above, no ~20 kDa crosslink product was obtained when cytosol was omitted (Figure 3A, lanes 9 and 10). When purified CaM (from bovine brain) and calcium were added, however, the ~20 kDa crosslink product was obtained (Figure 4, lanes 7 and 8), suggesting that the p-Prl SPF is crosslinked to CaM. The ~20 kDa crosslink product was also obtained when CaM prepared from *Dictyostelium discoideum* was added (Figure 4, lanes 9–12). CaM from *D.discoideum* was selected because a specific antiserum against this protein was available. With

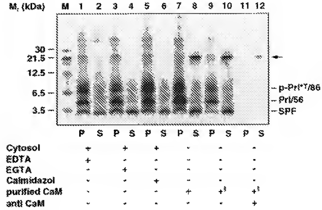


Fig. 4. Identification of CaM as the crosslink partner. p-Prl*^{T}/86 chains were released from the ribosome by EDTA (lanes 1 and 2) or puromycin (lanes 3–12) in the presence of cytosol (lanes 1–6), EGTA (lanes 3 and 4) or the CaM antagonist calmidazol (lanes 5 and 6). Cytosol was omitted in lanes 7–12 but purified CaM from bovine brain (lanes 7 and 8) or *D.discoideum* (§, lanes 9–12) were added instead. After UV irradiation membranes (P) were separated from soluble components (S) by centrifugation and analysed for crosslink products or immunoprecipitated with antibodies directed against *D.discoideum* CaM (lanes 11 and 12). Crosslinks to CaM are indicated by an arrow.

this antiserum we could immunoprecipitate the ~20 kDa crosslink product and thus further characterize its identity (Figure 4, lanes 11 and 12).

The p-Prl SPF is less efficiently released into the cytosol when factors are present that prevent an interaction with CaM (Figure 4, lanes 1–6) or when cytosol, and hence CaM, is absent (Figure 3A, lanes 9 and 10). This suggests that the interaction with CaM may facilitate the cytosolic localization of the amphipathic p-Prl SPF, which otherwise remains preferentially in the lipid bilayer.

The efficiency of crosslinking between the p-Prl SPF in the cytosol and CaM was very high, up to 55% (estimated from Figure 3A, lanes 2 and 4), and indicates that the majority of p-Prl SPF was in contact with CaM. Because the amount of p-Prl SPF generated in our *in vitro* translation/crosslinking system is very low (20–100 fmol/20 μl reaction), the high crosslinking efficiency also indicates a high affinity of CaM for the p-Prl SPF. Furthermore, the high crosslinking efficiency is consistent with a tight interaction between the p-Prl SPF and CaM [for chemical properties of the carbene-generating (Tmd)Phe see Brunner, 1989]. Similar crosslinking efficiencies have been reported, for example, for the tight interaction between the signal sequence of a growing polypeptide chain and the 54 kDa subunit of the signal recognition particle during protein targeting (High *et al.*, 1993b; Martoglio *et al.*, 1995).

Release of a SPF of HIV-1 p-gp160 into the cytosol and interaction with calmodulin

The characteristic feature of a CaM binding domain is a stretch of 16–35 amino acid residues that can potentially form a basic amphiphilic α(baa)-helix (James *et al.*, 1995). Such a stretch is predicted for the N-terminal portion of the p-Prl signal sequence. To see whether other signal sequences may also interact with CaM, we searched the signal sequences of mammalian and viral proteins listed

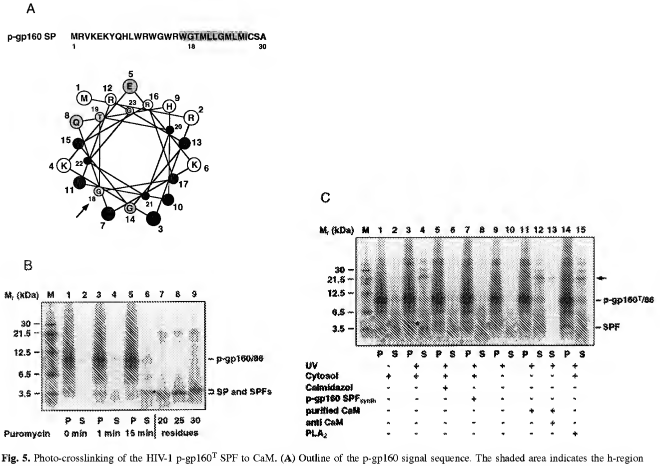


Fig. 5. Photo-crosslinking of the HIV-1 p-gp160^{T} SPF to CaM. (**A**) Outline of the p-gp160 signal sequence. The shaded area indicates the h-region of the signal sequence. The N-terminal 23 residues are also illustrated in a helical wheel; hydrophobic residues are indicated as dark circles, basic residues as white circles. The arrow indicates the amino acid (G18) that was replaced with (Tmd)Phe for site-specific photo-crosslinking. (**B**) Release of the p-gp160 SPF into the cytosol. p-gp160/86 chains were inserted into rough microsomes (lanes 1 and 2) and subsequently released from the ribosome by addition of puromycin. Incubation was continued for 1 (lanes 3 and 4) or 15 min (lanes 5 and 6) and membranes (pellet, P) separated from the cytosol (supernatant, S) by centrifugation before SDS–PAGE. Lanes 4–6 show *in vitro* synthesized peptides corresponding to the 20, 25 and 30 N-terminal amino acid residues of p-gp160. The p-gp160 SPF released into the cytosol is indicated by a dot (lane 6). (**C**) Photo-crosslinking of the p-gp160^{T} SPF to CaM. p-gp160^{T}/86 chains were released from the ribosome by puromycin in the presence of cytosol prepared from Jurkat T cells (lanes 1–8, 14 and 15). Where indicated, calmidazol (lanes 5 and 6) or synthetic p-gp160 SPF corresponding to the 23 N-terminal residues of the p-gp160 signal sequence (lanes 7 and 8) was added in addition. Cytosol was omitted in lanes 9–13 and purified CaM from *D.discoideum* and Ca^{2+} were added in lanes 11–13. Samples were incubated for 15 min at 22°C and subjected to UV light (lanes 3–15) and membranes (P) separated from soluble components (S) by centrifugation. Membranes and the cytosol fraction of one sample were then treated with phospholipase A_2 (lanes 14 and 15). Samples were finally analysed for crosslink products or immunoprecipitated with antibodies directed against *D.discoideum* CaM (lane 13) respectively. Crosslinks to CaM are indicated by an arrow, crosslinks to lipids by a star.

in the SWISSPROT database for their potential to form a baa-helix. Most signal sequences are short (<20 amino acid residues) and, after cleavage by signal peptidase and processing by signal peptide peptidase, are not expected to bind to CaM. However, we found one more signal sequence comprising 30 amino acid residues and consensus features for CaM binding. The signal sequence of the HIV-1 envelope protein p-gp160 has all the features for a CaM binding peptide (Figure 5A). The n-region of the HIV-1 envelope protein p-gp160 signal sequence can potentially form a baa-helix and contains several tryptophan residues often found in CaM binding domains (Vorherr *et al.*, 1990; James *et al.*, 1995).

With p-gp160 we performed analogous signal peptide release and crosslinking experiments as described above for p-Prl. Short p-gp160 chains (86 residues) were synthesized *in vitro* and inserted into the ER translocation sites of microsomal membranes (Figure 5B, lane 1). When p-gp160/86 chains were released from the ribosome by addition of puromycin and membranes were separated from the cytosol after 1 and 15 min incubation, a [^{35}S]-methionine-labelled peptide with an apparent molecular weight of 2–3 kDa appeared in the cytosol (Figure 5B, lanes 4 and 6). Because p-gp160/86 chains contain methionine residues only in the signal sequence, the released labelled peptide is either the cleaved signal sequence or a fragment thereof.

To determine the approximate length of the peptide released into the cytosol, we synthesized marker peptides representing the entire p-gp160 signal sequence (30 amino acid residues) or N-terminal SPFs of 25 and 20 amino acid residues respectively. Comparative analysis of the peptides separated by SDS–PAGE revealed an estimated size of 20–25 amino acid residues, clearly smaller than the entire signal sequence (Figure 5B, lanes 6–9). This indicated that the released peptide is a fragment of the p-gp160 signal sequence and suggests that the p-gp160 signal sequence is rapidly processed. We could not detect

a peptide corresponding to the entire signal sequence, as was the case for the signal sequence of p-Prl. The predicted processing site of the p-Prl signal peptide is between the two leucine clusters (Figure 1B) of its h-region (Lyko *et al.*, 1995). Whether such a motif is required for signal sequence processing is not known. The h-region of the p-gp160 signal sequence also contains two clusters of amino acids with long hydrophobic side chains (-MLLGMLMI-) between which processing may occur (see Figure 5A).

We next probed the molecular environment of the released p-gp160 SPF using site-specific photo-crosslinking. The photo-activatable amino acid (Tmd)Phe was co-translationally incorporated into the p-gp160 signal sequence instead of Gly18 (see Figure 5A) to give p-gp160^T and p-gp160^T/86 chains which were used for membrane insertion and puromycin release as described above. Crosslinking was induced with UV light 15 min after addition of puromycin. We found the SPF and a major crosslink product with an apparent molecular weight of ~20 kDa in the cytosol fraction (Figure 5C, lane 4, arrow). This indicates that the peptide released into the cytosol contains (Tmd)Phe and hence must be derived from the p-gp160 signal sequence. The cytosol used for these experiments was prepared from Jurkat T cells. The same results were also obtained when cytosol prepared from bovine brain was used (not shown).

To test whether the p-gp160^T SPF is crosslinked to CaM, we released p-gp160^T/86 chains from the ribosome in the presence of the CaM antagonist calmidazol. Furthermore, we released p-gp160^T/86 chains when cytosol was omitted and when purified CaM and Ca^{2+} were added instead. The ~20 kDa crosslink product was not observed in the presence of calmidazol (Figure 5C, lanes 5 and 6) or when cytosol was omitted (Figure 5C, lanes 9 and 10). The ~20 kDa crosslink product was obtained, however, when calcium and purified CaM from *D.discoideum* (Figure 5C, lanes 11–12) or bovine brain (not shown) were present. In addition, we could immunoprecipitate the ~20 kDa crosslink product with antiserum against CaM from *D.discoideum* (Figure 5C, lane 13). These results indicate that the p-gp160^T SPF interacts with CaM. In analogy to p-Prl and based on the consensus for CaM binding, we expect that the released p-gp160^T SPF comprises the N-teminal part of the signal sequence.

Not all the p-gp160 SPF was released into the cytosol. After crosslinking a low molecular weight crosslink product appeared in the membrane fractions (Figure 5C, lane 3, star). This 4–5 kDa crosslink product was sensitive to phospholipase A_2, indicating that it is a phospholipid adduct (Figure 5C, lane 14). Thus, some p-gp160^T SPF remains in the membrane in contact with phospholipids.

Characterization of the p-gp160 SPF–CaM complex

CaM interacts with target proteins with affinity constants in the low nanomolar range (James *et al.*, 1995). We determined the affinity constant for formation of the p-gp160 signal peptide–CaM complex by fluorometric titration using dansylated CaM (Anderson and Malencik, 1986) and a synthetic peptide corresponding to the 23 N-terminal residues of the p-gp160 signal sequence. This peptide could efficiently prevent crosslinking between the p-gp160^T SPF and CaM in the signal peptide release and crosslinking experiment with p-gp160^T/86 chains (Figure 5C, lanes 7 and 8). In the presence of calcium, dansyl-CaM showed a large increase in the fluorescence intensity upon binding of the peptide (Figure 6A). The K_d determined from two series of three titrations was 22 $\pm$ 5 nM and was derived from a non-linear curve fitting procedure. No fluorescence enhancement was observed when samples contained EGTA and no calcium (not shown).

We also determined the peptide–CaM interaction in an enzyme inhibition experiment. The inhibitory effect of the synthetic p-gp160 SPF on purified CaM-dependent cyclic nucleotide phosphodiesterase (PDE) was tested *in vitro* (Wallace *et al.*, 1984). The result of such an experiment is expressed as the concentration that inhibits enzymatic activity by 50% (IC_{50}) in solutions initially containing sufficient CaM to produce 50% of the maximum stimulation (~8 nM CaM; Figure 6B, upper panel) (Anderson and Malencik, 1986). For the p-gp160 SPF we determined an IC_{50} value of ~30 nM (Figure 6B, lower panel).

We could not determine the K_d and IC_{50} values for the p-Prl SPF–CaM complex because synthesis of the p-Prl SPF was technically not feasible due to the high number of leucine residues. However, similar high crosslinking efficiencies of p-gp160^T and p-Prl*^{T} SPFs with CaM indicate similar affinities for CaM.

Discussion

We have identified components that interact with the signal peptide of p-Prl in the ER membrane and with a SPF comprising the N-terminus of the p-Prl signal peptide in the cytosol. Right after cleavage by signal peptidase the p-Prl signal sequence accumulates in the membrane and was found to be in contact with lipids as well as the 18 and 21 kDa subunits of SPC. These two SPC subunits share homology with the *Escherichia coli* leader peptidase and the yeast Sec11 protein, shown to be an essential component of the yeast signal peptidase complex (Greenburg *et al.*, 1989; van Dijl *et al.*, 1992). We have previously shown that the p-Prl signal sequence before cleavage is in contact with lipids as well as components of the protein-conducting channel (TRAMp and Sec61α) (Martoglio *et al.*, 1995). Our present data suggest that after cleavage by signal peptidase the p-Prl signal sequence moves from the translocon–lipid interface into proximity to SPC18 and SPC21, the putative catalytic subunits of SPC. However, no direct evidence for such a function has been obtained for the two SPC subunits.

Previously it has been shown that the cleaved p-Prl signal sequence is rapidly processed in the ER membrane and an N-terminal SPF is released into the cytosol (Lyko *et al.*, 1995). We show here with site-specific photo-crosslinking that the released p-Prl SPF can be efficiently crosslinked to CaM in a Ca^{2+}-dependent manner. Activation of (Tmd)Phe with UV light generates a highly reactive carbene ($t_{1/2}$ ~ 1 ns) which, in turn, reacts with any adjacent molecule, independent of whether it is a protein, a lipid or a water molecule (Brunner, 1989). Due to this unique property of the carbene, we can conclude from the crosslinking efficiency that in our cell-free system the released p-Prl SPF has high affinity for CaM. Whether such conditions are also found in cells synthesizing Prl remains to be established.

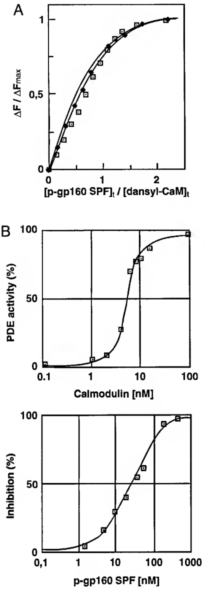


Fig. 6. Characterization of the p-gp160 SPF–CaM complex. (**A**) Fluorescence titration of dansyl-CaM with p-gp160 SPF. Dansylated CaM [160 (◆) or 260 nM (□)] was titrated with a synthetic peptide corresponding to the 23 N-terminal residues of the p-gp160 signal sequence. The fluorescence intensity at 470 nm after excitation at 340 nm was recorded at each concentration of peptide. The fluorescence enhancement values were normalized for maximal enhancement at saturation and the values of three independent experiments were plotted against the ratio peptide:dansyl-CaM. (**B**) Effect of the p-gp160 SPF on CaM-dependent PDE-mediated hydrolysis of cAMP. Activator-deficient PDE was preincubated in the assay medium with increasing CaM concentrations (upper panel) or in the presence of 8 nM CaM and increasing concentrations of synthetic p-gp160 SPF (lower panel). The enzyme reaction was started by addition of cAMP. CaM-induced PDE activity is plotted against CaM concentration (upper panel), showing half maximal activation of PDE at 8 nM CaM. Inhibition of PDE activity at 8 nM CaM is plotted against p-gp160 SPF concentration (lower panel). The determined IC_{50} is ~30 nM.

CaM recognizes positively charged, amphiphilic α-helical stretches of 16–35 amino acid residues in CaM binding proteins as well as peptides mimicking such sites (O'Neil and DeGrado, 1990; James *et al.*, 1995). The p-Prl signal sequence has an extended n-region and the p-Prl SPF can potentially form a basic amphiphilic α-helix. We also found this feature in the signal sequence of the HIV-1 envelope protein p-gp160, but not in the majority of other signal sequences. Thus, only a few signal sequences would yield SPFs that efficiently interact with CaM after proteolytic processing.

Cleavage of the p-gp160 signal sequence from p-gp160/86 was very inefficient in our cell-free system. It is reported that signal sequence cleavage of p-gp160 is inefficient *in vivo*, which may indicate an intrinsic property of p-gp160 (Li *et al.*, 1996). The limiting amount of p-gp160 SPF released into the cytosol did not allow rigorous charcterization of the p-gp160 SPF, as shown for the p-Prl SPF. However, the size of the peptide, the presence of radioactive label and (Tmd)Phe as well as crosslinking to CaM indicate that the peptide released into the cytosol is a fragment of the p-gp160 signal sequence and contains features for CaM binding. The p-Prl SPF has been shown to comprise the N-terminal part of the signal sequence. By analogy with these data, it is likely that the p-gp160 SPF released into the cytosol also corresponds to the N-terminal part of the signal sequence. A synthetic peptide corresponding to the 23 N-terminal residues of the p-gp160 signal sequence shows low K_d and IC_{50} values, indicating high affinity of this SPF for CaM.

p-Prl and p-gp160 SPFs may act as CaM antagonists

What could be the functional and physiological significance of an interaction between the p-Prl and p-gp160 SPFs and CaM? In principle the SPF–CaM complex could acquire a novel function similar to activation of MHC molecules by peptides (Heemels and Ploegh, 1995). Another possibility is that SPFs function as CaM antagonists in the vicinity of the rough ER membrane. CaM-dependent processes may be inhibited when large amounts of a CaM binding SPF are generated and released into the cytosol. The obvious question for the latter role is whether the local concentration of SPFs can be high enough to impair CaM functions. CaM is a highly abundant protein and can account for 0.2–12 μg (11–700 pmol)/mg total cell protein depending on the cell type (Klee and Vanaman, 1982). The free CaM concentration in the cytosol, however, is difficult to estimate. CaM is often associated with CaM-dependent enzymes and considerable differences in the subcellular localization of CaM have been reported (Klee and Vanaman, 1982). The ER membrane is not especially rich in CaM-dependent enzymes and does not show an accumulation of CaM at its surface (D.Guerini, personal communication). Thus the CaM concentration at the ER surface may be similar to the free CaM concentration, which is thought to be in the range 10–100 μM (Klee and Vanaman, 1982; D.Guerini, personal communication).

In a stimulated rat pituitary cell line the amount of Prl synthesized within 10 min is ~60 ng (~3 pmol)/mg total cell protein (Gordeladze, 1990). Considering that the cytosol contains ~40% of the total cell protein and the protein concentration in the cytosol is 200–300 mg/ml (Alberts *et al.*, 1994), the p-Prl SPF concentration in the cytosol can rapidly reach ~2 μM (200–300 mg/ml×100/40×3 pmol/mg), provided that all the p-Prl SPF is released

into the cytosol. Because the SPF is released only from the rough ER, the local SPF concentration is certainly higher and would exceed the local CaM concentration.

It is also likely that high levels of p-gp160 SPF will be generated when gp160 is expressed in HIV-1-infected cells. For example, the steady-state level of gp160 expressed in chronically infected cell lines has been shown to exceed the levels of actin, a highly abundant protein in mammalian cells (Geleziunas *et al.*, 1994). We therefore can expect micromolar concentrations of p-gp160 SPF in the cytosol and much higher local concentrations at the ER surface.

Release of SPFs from the ER membrane appeares to be a regulated process. It was shown that the proline isomerase inhibitor cyclosporin A inhibits processing of the cleaved signal sequence (Klappa *et al.*, 1996). This suggests that a cyclophilin may modulate activity of the signal peptide peptidase and thereby regulate release of the SPF into the cytosol. Thus, a burst of SPFs may be released from the ER membrane, similarly to Ca^{2+} released from the ER lumen, and may transiently interfere with CaM-dependent processes.

The role of CaM in the regulation of cellular processes has been investigated in many studies. In experiments with tissue culture cells CaM antagonists were shown to inhibit CaM-dependent processes at low micromolar concentrations (White, 1985; Srinivas *et al.*, 1994; de Figueiredo and Brown, 1995). CaM antagonists used in these experiments (calmidazol, trifluoperazine, W13 and W7) have similar or several fold higher IC_{50} values for PDE (0.04–68 μM) than the synthetic p-gp160 SPF used in the present study. The experiments show that under physiological conditions CaM antagonists can severely affect CaM function at low micromolar concentration even though the estimated intracellular CaM concentration is considerably higher.

Possible functions of the p-Prl and p-gp160 SPFs

A possible function for the p-Prl SPF can be envisaged in Prl secretion. cAMP plays a central role in Prl expression and secretion in the anterior pituitary (Lamberts and MacLeod, 1990). The intracellular level of cAMP is regulated by adenylate cyclases (AC), synthesizing cAMP, and PDEs, hydrolysing cAMP. Ca^{2+}/CaM-dependent types of both enzymes have been characterized in mammalian cells. In the rat anterior pituitary, cells that predominantly produce Prl do not express Ca^{2+}/CaM-dependent ACs but Ca^{2+}/CaM-dependent PDE (Gordeladze, 1990; Paulssen *et al.*, 1994). The p-Prl SPF may thus compete with Ca^{2+}/CaM-dependent PDE for CaM and inhibit PDE. This would lead to prolonged cAMP signalling and continued stimulation of Prl secretion.

Viruses use many strategies to escape immune detection. They can interfere with antigen presentation or block signal transduction pathways. In HIV-1 infected cells several CaM-dependent processes involved in immune defence are disrupted (Miller *et al.*, 1993). One mechanism by which this may be achieved is binding of CaM to the cytoplasmic domain of gp41 protein (Miller *et al.*, 1993; Srinivas *et al.*, 1993). Additional binding of CaM by the p-gp160 SPF, as we show here, may further contribute to inactivation of CaM-dependent processes in infected T cells (Srinivas *et al.*, 1994).

Our results imply that signal sequences of distinct secretory proteins and membrane proteins may have a function in addition to protein targeting and translocation. The signal sequences of p-Prl and p-gp160 have extended n-regions (>10 residues), in contrast to most other signal sequences. Interestingly, the gene for Prl encodes the signal sequence by two exons separating the n- and h-regions. In many cases exons encode distinct domains or functional regions of the final protein (Gilbert, 1985). The first exon of the *Prl* gene codes for 10 amino acid residues providing the hydrophilic and basic residues, an essential CaM binding feature, to the signal sequence of p-Prl (Troung *et al.*, 1984). Further studies with mutant p-Prl signal sequences will be required to show whether the amino acids encoded by exon 1 confer on this signal sequence an additional function and to exclude other explanations for this splicing junction.

Besides CaM, other cytosolic and even nuclear proteins may be targets for SPFs. SPFs may be derived from signal sequences with exceptionally long n-regions. Such signal sequences are often found on viral membrane proteins, like those from lentiviruses or LCMV (Pancino *et al.*, 1994; Hombach *et al.*, 1995). The length and high conservation of lentiviral signal sequences has already prompted speculations about alternative functions. Our results indicate an additional function for one type of signal sequence and demonstrate, in addition, a convenient way to identify targets for SPFs in general. Besides identifying new targets of SPFs, physiological studies are now required to determine the cellular effects of SPFs in their native cellular and organism contexts.

Materials and methods

Materials

General chemicals were from Merck (Darmstadt, Germany) or Sigma (München, Germany). Restriction enzymes, SP6 RNA polymerase and bee venom phospholipase A_2 were from Boehringer Mannheim (Mannheim, Germany). Activator-deficient PDE, 5′-nucleotidase and Fiske–SubbaRow reducer were from Sigma (München, Germany). Vectors pGEM3Z and pGEM4Z, as well as reticulocyte lysate, were from Promega (Heidelberg, Germany). [^{35}S]methinonine was from Amersham Buchler (Braunschweig, Germany). Calmidazol and calmodulin were from Calbiochem (La Jolla, CA). Plasmid pL102 coding for p-gp160 was kindly provided by V.Bosch (DKFZ, Heidelberg, Germany). *Dictyostelium discoideum* CaM and the corresponding antiserum were generously provided by T.Soldati and B.Ulbricht (MPI, Heidelberg, Germany). The antibodies against SPC12, SPC18, SPC21 and SPC22/23 were gifts from E.Hartmann (MDC, Berlin, Germany) and C.Nicchitta (Duke University, Durham, NC).

Plasmids and transcription

Plasmid pL102 coding for p-gp160 was digested with *Sal*I and *Xho*I and the insert DNA transferred into the *Sal*I site of pGEM4Z under control of the SP6 promotor. The plasmid encoding p-Prl* (pGEM4Z/p-Prl*) has been described previously (Lyko *et al.*, 1995). The coding region of p-Prl* was transferred into pGEM3Z under control of the SP6 promotor to give pGEM3Z/p-Prl*. Codon 18 of the coding regions of p-Prl* and p-gp160 were replaced by the TAG codon using overlap extension PCR (Ho *et al.*, 1989) to give pGEM3Z/p-Prl*^{T} and pGEM4Z/p-gp160^T. To prepare mRNA coding for p-Prl*/86 and p-Prl*^{T}/86, the respective plasmids were linearized with *Pvu*II and transcribed with SP6 RNA polymerase (Lyko *et al.*, 1995). To prepare mRNA coding for p-gp160*^{T}/86, the respective coding region was amplified by PCR and transcribed with SP6 RNA polymerase (Nilsson and von Heijne, 1993).

Translation and photo-crosslinking

Truncated transcripts encoding the 86 N-terminal amino acids of p-Prl*, p-Prl*^{T} and p-gp160*^{T} were translated for 15 min at 30°C in 12.5 μl reticulocyte lysate containing [^{35}S]methinonine, suppressor tRNA, SRP

and dog pancreatic rough microsomes (Martoglio *et al.*, 1995). After translation, the salt concentration was increased to 500 mM KOAc and the samples incubated for 5 min on ice. Membranes were then separated by a 3 min centrifugation through a 150 µl sucrose cushion at 48 000 r.p.m. and 4°C in a Beckman TLA100 rotor (Lyko *et al.*, 1995). The tube containing the membrane pellet was carefully rinsed with 200 µl RM buffer [50 mM HEPES–KOH, pH 7.6, 100 mM KOAc, 3 mM $Mg(OAc)_2$ and 2 mM dithiothreitol]. The membrane pellet was resuspended in 10 µl cytosol and 10 µl RM buffer. Cytosol was prepared from bovine brain (Celis, 1994), GH_3 cells, Mel Juse cells or Jurkat cells (Smythe *et al.*, 1992). Wheatgerm extract was prepared according to Erickson and Blobel (1983). Where indicated, EDTA (25 mM), EGTA (5 mM), calmidazol (20 µM), CaM from bovine brain or *D.discoideum* (200 nM) and calcium chloride (100 µM) or synthetic p-gp160 SPF (20 µM) were also added. When cytosol was omitted, membranes were resuspended in 20 µl RM buffer. Nascent chains were released by adding 0.8 µl 100 mM puromycin and incubation at 22°C for 1 or 15 min. For crosslinking, samples were put on ice and UV irradiated (364 nm) for 2 min (Martoglio *et al.*, 1995). Membranes were separated from the cytosol by a 10 min centrifugation at 100 000 r.p.m. and 2°C in a Beckman TLA100 rotor.

Analysis of translation and crosslink products

Translation and crosslink products were analysed in 16.5% T, 6% C polyacrylamide gels according to Schägger and von Jagow (Schägger and von Jagow, 1987; Lyko *et al.*, 1995). Labelled proteins were visualized with a Fuji phosphorimager BAS1000. For immunoprecipitation, proteins were denatured in 1% SDS, solubilized in IP buffer (10 mM Tris–HCl, pH 7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100) and incubated with the relevant antibodies (High *et al.*, 1993a). Antibodies were raised against peptides corresponding to the 14 N-terminal amino acids of the p-Prl signal sequence (anti-p-Prl SPF) and to residues 38–50 of Prl (anti-Prl). For phospholipid analysis membranes were resuspended in 20 µl RM buffer with 5 mM $CaCl_2$ and treated with 10 U bee venom phospholipase A_2 for 5 min at 41°C (Martoglio *et al.*, 1995).

Fluorescence measurements

CaM binding by a synthetic peptide corresponding to the N-terminal 23 residues of the p-gp160 signal sequence was determined by fluorometry using dansyl-CaM according to published procedures (Anderson and Malencik, 1986; Vorherr *et al.*, 1990). The fluorescence emission of dansyl-CaM in the absence and presence of peptide was scanned from 400 to 550 nm after excitation at 340 nm using a Shimadzu RF-5000 spectrofluorometer. Samples (3 ml) contained 50 mM Tris–HCl, pH 7.5, 150 mM KCl, 1 mM $CaCl_2$ or 1 mM EGTA, 160 or 260 nM dansyl-CaM (Vorherr *et al.*, 1990) and increasing concentrations of peptide. The dissociation constant was calculated by a non-linear curve fitting procedure using Enzfitter software (Elsevier-Biosoft, Cambridge).

Enzyme inhibition assay

To establish a calibration curve of PDE activity in the presence of CaM, 0.004 U activator-deficient PDE were preincubated for 10 min at 30°C in 100 µl 50 mM Tris–HCl, pH 8.0, 5 mM $MgSO_4$, 100 µM $CaCl_2$, 0.02% Triton X-100 and various concentrations of CaM (Wallace *et al.*, 1984). The reaction was started by addition of 2 µl 100 mM cAMP. After incubation for 30 min at 30°C the reaction was terminated by boiling the sample for 2 min at 95°C. 5′-AMP was subsequently converted to adenosine with 5′-nucleotidase and the phosphate released was measured by the method of Fiske and SubbaRow (Wang and Desai, 1977). Inhibition assays were performed as above except that PDE was preincubated with 8 nM CaM (required for 50% CaM-induced PDE activity as determined from the calibration curve) and various concentrations of synthetic p-gp160 SPF.

Acknowledgements

This work is written in memory of Roland Graf. We thank T.Soldati and B.Ulbricht for *D.discoideum* calmodulin and the corresponding antiserum, E.Hartmann and C.Nicchitta for anti-SPC antibodies and V.Bosch for plasmid pL102. Special thanks are due to J.Brunner for the synthesis and purification of (Tmd)Phe. Many thanks are also due to I.Braakman, D.Guerini, G.Bacher and O.Gruss for stimulating discussions and critical comments. This work was supported by grants from the DFG and the Fonds der Chemischen Industrie. R.G. was supported by the Swiss National Science Foundation (grant to J.Brunner).

References

Alberts,B., Bray,D., Levis,J., Raff,M., Roberts,K. and Watson,J. (1994) *Molecular Biology of the Cell.* 3rd edn. Garland Press, New York.

Anderson,S.R. and Malencik,D.A. (1986) Peptides recognizing calmodulin. *Calmodulin Cell Function*, **6**, 2–41.

Blobel,G. and Dobberstein,B. (1975) Transfer of proteins across membranes II. Reconstitution of functional rough microsomes from heterologous components. *J. Cell Biol.*, **67**, 852–862.

Böhni,P.C., Deshaies,R.J. and Schekman,R. (1988) SEC11 is required for signal peptide processing and yeast cell growth. *J. Cell Biol.*, **106**, 1035–1042.

Brunner,J. (1989) Photochemical labelling of apolar phase of membranes. *Methods Enzymol.*, **172**, 628–687.

Celis,J.E. (1994) *Cell Biology: A Laboratory Handbook.* 1st edn. Academic Press, San Diego, CA.

de Figueiredo,P. and Brown,W.J. (1995) A role for calmodulin in organelle membrane tubulation. *Mol. Biol. Cell*, **6**, 871–887.

Erickson,A.H. and Blobel,G. (1983) Cell-free translation of messenger RNA in a wheat germ system. *Methods Enzymol.*, **96**, 38–50.

Evans,E.A., Gilmore,R. and Blobel,G. (1986) Purification of microsomal signal peptidase as a complex. *Proc. Natl Acad. Sci. USA*, **83**, 581–585.

Geleziunas,R., Bour,S. and Wainberg,M.A. (1994) Correlation between high level gp160 expression and reduced CD4 biosynthesis in clonal derivatives of human immunodeficiency virus type 1-infected U-937 cells. *J. Gen. Virol.*, **75**, 857–865.

Gilbert,W. (1985) Genes-in-pieces revisited. *Science*, **228**, 823–824.

Gilmore,R., Collins,P., Johnson,J., Kellaris,K. and Rapiejko,P. (1991) Transcription of full-length and truncated mRNA transcripts to study protein translocation across the endoplasmic reticulum. *Methods Cell Biol.*, **34**, 223–239.

Gordeladze,J.O. (1990) The pharmacodynamic action of the cyclic AMP phosphodiesterase inhibitor Rolipram on prolactin producing rat pituitary adenoma (GH_4C_1) cells. *Biosci. Rep.*, **10**, 375–388.

Greenburg,G., Shelness,G.S. and Blobel,G. (1989) A subunit of mammalian signal peptidase is homologous to yeast SEC11 protein. *J. Biol. Chem.*, **264**, 15762–15765.

Heemels,M.T. and Ploegh,H. (1995) Generation, translocation, and presentation of MHC class I-restricted peptides. *Annu. Rev. Biochem.*, **64**, 463–491.

Henderson,R.A., Michel,H., Sakaguchi,K., Shabanowitz,J., Appella,E., Hunt,D. and Engelhard,V.H. (1992) HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. *Science*, **255**, 1264–1266.

High,S., Andersen,S.S.L., Görlich,D., Hartmann,E., Prehn,S., Rapoport, T.A. and Dobberstein,B. (1993a) Sec61p is adjacent to nascent type I and type II signal-anchor proteins during their membrane insertion. *J. Cell Biol.*, **121**, 743–750.

High,S. *et al.* (1993b) Site-specific photocross-linking reveals that Sec61p and TRAM contact different regions of a membrane inserted signal sequence. *J. Biol. Chem.*, **268**, 26745–26751.

Ho,S.N., Hunt,H.D., Horton,R.M., Pullen,J.K. and Pease,L.R. (1989) Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, **77**, 51–59.

Hombach,J., Pircher,H., Tonegawa,S. and Zinkernagel,R.M. (1995) Strictly transporter of antigen presentation (TAP)-dependent presentation of an immunodominant cytotoxic T lymphocyte epitope in the signal sequence of a virus protein. *J. Exp. Med.*, **182**, 1615–1619.

James,P., Vorherr,T. and Carafoli,E. (1995) Calmodulin-binding domains: just two faced or multi-faceted? *Trends Biochem. Sci.*, **20**, 38–42.

Klappa,P., Dierks,T. and Zimmermann,R. (1996) Cyclosporin A inhibits the degradation of signal sequences after processing of presecretory proteins by signal peptidase. *Eur. J. Biochem.*, **239**, 509–518.

Klee,C.B. and Vanaman,T.C. (1982) Calmodulin. *Adv. Protein Chem.*, **35**, 213–321.

Lamberts,S.W.J. and MacLeod,R.M. (1990) Regulation of prolactin secretion at the level of the lactotroph. *Physiol. Rev.*, **70**, 279–325.

Li,Y., Bergeron,J.J.M., Luo,L., Ou,W.J., Thomas,D.Y. and Kang,Y. (1996) Effects of inefficient cleavage of the signal sequence of HIV-1 gp120 on its association with calnexin, folding and intracellular transport. *Proc. Natl Acad. Sci. USA*, **93**, 9606–9611.

Lyko,F., Martoglio,B., Jungnickel,B., Rapoport,T.A. and Dobberstein,B. (1995) Signal sequence processing in rough microsomes. *J. Biol. Chem.*, **270**, 19873–19878.

Martoglio,B. and Dobberstein,B. (1996) Snapshots of membrane translocating proteins. *Trends Cell Biol.*, **6**, 142–147.

Martoglio,B., Hofmann,M., Brunner,J. and Dobberstein,B. (1995) The protein conducting channel in the membrane of the endoplasmic reticulum is open laterally toward the lipid bilayer. *Cell*, **81**, 207–214.
Miller,M.A., Mietzner,T.A., Cloyd,M.W., Robey,G. and Montelaro,R.C. (1993) Identification of a calmodulin-binding and inhibitory peptide domain in the HIV-1 transmembrane glycoprotein. *AIDS Res. Hum. Retrovir.*, **9**, 1057–1066.
Nilsson,I. and von Heijne,G. (1993) Determination of the distance between oligossaccharyltransferase active site and the endoplasmic reticulum membrane. *J. Biol. Chem.*, **268**, 5798–5801.
O'Neil,K.T. and DeGrado,W.F. (1990) How calmodulin binds its targets: sequence independent recognition of amphipathic α-helices. *Trends Biochem. Sci.*, **15**, 59–64.
Pancino,G., Ellerbrok,H., Sitbon,M. and Sonigo,P. (1994) Conserved framework of envelope glycoproteins among lentiviruses. *Curr. Topics Microbiol. Immunol.*, **188**, 77–105.
Paulssen,R.H., Johansen,P.W., Gordeladze,J.O., Nymoen,O., Paulssen, E.J. and Gautvik,K.M. (1994) Cell-specific expression and function of adenylyl cylclases in rat pituitary tumor cell lines. *Eur. J. Biochem.*, **222**, 97–103.
Rapoport,T., Jungnickel,B. and Kutay,U. (1996) Protein translocation across the eukaryotic endoplsmic reticulum and bacterial inner membranes. *Annu. Rev. Biochem.*, **65**, 271–303.
Schägger,H. and von Jagow,G. (1987) Tricine–sodium dodecyl sulfate–polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal. Biochem.*, **166**, 368–379.
Schreiber,S.L. and Crabtree,G.R. (1992) The mechanism of action of cyclosporin A and FK506. *Immunol. Today*, **13**, 136–142.
Shelness,G.S. and Blobel,G. (1990) Two subunits of canine signal peptidase complex are homologous to yeast SEC11 protein. *J. Biol. Chem.*, **265**, 9512–9519.
Smythe,M., Redelmeier,T.E. and Schmid,S.L. (1992) Receptor-mediated endocytosis in semiintact cells. *Methods Enzymol.*, **219**, 223–234.
Srinivas,R.V., Bernstein,H., Oliver,C. and Compans,R.W. (1994) Calmodulin antagonists inhibit human immunodeficiency virus-induced cell fusion but not virus replication. *AIDS Res. Hum. Retrovir.*, **10**, 1489–1496.
Srinivas,S.K., Srinivas,R.V., Anantharamaiah,G.M., Compans,R.W. and Segrest,J.P. (1993) Cytosolic domain of the human immunodeficiency virus envelope glycoprotein binds to calmodulin and inhibits calmodulin regulated proteins. *J. Biol. Chem.*, **268**, 22895–22899.
Troung,A.T., Duez,C., Belayew,A., Renard,A., Pictet,R., Bell,G.I. and Martial,J.A. (1984) Isolation and characterization of the human prolactin gene. *EMBO J.*, **3**, 429–437.
van Dijl,J.M., de Jong,A., Vehmaanperä,J., Venema,G. and Bron,S. (1992) Signal peptidase I of *Bacillus subtilis*: patterns of conserved amino acids in prokaryotic and eukaryotic type I signal peptidases. *EMBO J.*, **11**, 2819–2828.
von Heijne,G. (1985) Signal sequences. The limits of variation. *J. Mol. Biol.*, **184**, 99–105.
Vorherr,T., James,P., Krebs,J., Enyedi,A., McCormick,D.J., Pennistone, J.T. and Carafoli,E. (1990) Interaction of calmodulin with the calmodulin binding domain of the plasma membrane Ca^{2+} pump. *Biochemistry*, **29**, 355–365.
Wallace,R.W., Tallant,E.A. and Cheung,W.Y. (1984) Assay of calmodulin by Ca^{2+}-dependent phosphodiesterase. *Methods Enzymol.*, **102**, 39–47.
Wang,J.H. and Desai,R. (1977) Modulator binding protein. *J. Biol. Chem.*, **252**, 4175–4184.
Wei,M.L. and Cresswell,P. (1992) HLA-A2 molecules in an antigen processing mutant cell contain signal sequence-derived peptides. *Nature*, **356**, 443–446.
White,B. (1985) Evidence for a role of calmodulin in the regulation of prolactin gene expression. *J. Biol. Chem.*, **260**, 1213–1217.

Received on March 6, 1997; revised on August 27, 1997